Impact of chat layout on usability in customer service chat multitasking

Jenni Pajukoski

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 17.09.2018

Supervisor

Prof. Antti Oulasvirta

Advisors

Ph.D. Jussi Jokinen

M.Sc.(Tech.) Kimmo Kiiski

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Jenni Pajukoski

Title Impact of chat layout on usability in customer service chat multitasking

Degree programme Master's Programme in Computer, Communication and Information Sciences

Major Computer Science **Code of major** SCI3042

Supervisor Prof. Antti Oulasvirta

Advisors Ph.D. Jussi Jokinen, M.Sc.(Tech.) Kimmo Kiiski

Date 17.09.2018 **Number of pages** 54+24 **Language** English

Abstract

Over the last few years, chat has become an important channel in customer service. The demand for chat as a customer service channel is constantly growing, as more and more customers require real time help during their visits on companies' websites. This demand highlights the importance of chat agents' ability to handle multiple simultaneous chats at a time. However, it has been proven by numerous studies, that multitasking has decreasing effects on performance. On the other hand, studies have shown that those effects can be decreased with different user interface design decisions.

This thesis studies how user interface layout of a customer service chat system affects usability in chat multitasking. The research problem is addressed by conducting an experiment, where two popular customer service chat layouts are compared: In the first, windowed layout, all simultaneous chat windows are shown for the user at a time. In the second, tabbed layout, only one conversation is shown for the user at a time. In addition, differences between having either three or four simultaneous chats are investigated.

The results indicate that the windowed layout was faster in terms of efficiency, when first response time was measured. Other aspects of efficiency, measured as question response time and chat duration, did not differ between the layouts. However, when the chat amount was increased from three to four, it was found that the stress level increased considerably in the windowed layout. All aspects of efficiency decreased in both layouts when the amount of simultaneous chats was increased from three to four. However, increasing the chat amount did not affect any other satisfaction measures than the perceived stress level.

Because the response time for actual questions or preference between the layouts did not differ, both layouts can be suggested to be used in chat agent user interface. Slightly better choice would be the windowed layout, if first responses are wanted to be fast. However, the amount of visible chat windows should be considered carefully, to avoid increase in stress levels.

Keywords customer service chat, multitasking, usability, experiment

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Jenni Pajukoski

Työn nimi Chat-layoutin vaikutus käytettävyyteen asiakaspalvelussa usean samanaikaisen keskustelun aikana

Koulutusohjelma Master's Programme in Computer, Communication and Information Sciences

Pääaine Computer Science **Pääaineen koodi** SCI3042

Työn valvoja Prof. Antti Oulasvirta

Työn ohjaajat FT Jussi Jokinen, DI Kimmo Kiiski

Päivämäärä 17.09.2018 **Sivumäärä** 54+24 **Kieli** Englanti

Tiivistelmä

Muutaman viime vuoden aikana chatista on tullut tärkeä kanava asiakaspalvelussa. Tarve chatille asiakaspalvelukanavana kasvaa jatkuvasti, kun yhä useammat asiakkaat vaativat reaaliaikaista apua vieraillessaan yritysten nettisivuilla. Tämä tarve korostaa usean yhtäaikaisen chatin käsittelyn kyvyn tärkeyttä. Monet tutkimukset ovat kuitenkin osoittaneet, että usean asian tekeminen samaan aikaan laskee suorituskykyä. Toisaalta, tutkimukset ovat osoittaneet, että suorituskyvyn laskua voidaan ehkäistä erilaisilla käyttöliittymäratkaisuilla.

Tämä diplomityö tutkii asiakaspalvelu-chat-järjestelmän käyttöliittymän layoutin vaikutusta käytettävyyteen usean samanaikaisen keskustelun aikana. Tutkimusongelmaa tarkastellaan kokeella, jossa vertaillaan kahta yleisesti käytössä olevaa asiakaspalvelu-chat-layoutia: Ensimmäisessä layoutissa kaikki käynnissä olevat keskusteluikkunat näytetään käyttäjälle samaan aikaan. Toisessa layoutissa käyttäjälle näytetään vain yksi keskustelu kerrallaan. Kokeessa tutkitaan myös eroja kolmen ja neljän samanaikaisen chat-keskustelun välillä.

Tulosten perusteella vastausaika ensimmäiseen viestiin oli nopeampi ensimmäisellä layoutilla. Muut tehokkuuden mittarit eivät eronneet layoutien välillä. Kun samanaikaisten keskustelujen määrää kasvatettiin kolmesta neljään, stressin määrä kasvoi kuitenkin merkittävästi ensimmäisessä layoutissa, mutta ei toisessa layoutissa. Kaikki tehokkuuden mittarit laskivat, kun samanaikaisten keskustelujen määrä nostettiin kolmesta neljään. Samanaikaisten keskustelujen määrä ei kuitenkaan vaikuttanut tyytyväisyyden mittareihin muiden kuin stressin osalta.

Koska vastausaika varsinaisiin kysymyksiin ei eronnut layoutien välillä, ja yhtä moni ihminen piti enemmän ensimmäisestä kuin toisestakin layoutista, kumpaakin layoutia voidaan suositella käytettäväksi chat-asiakaspalvelijan käyttöliittymässä. Jos vastausajan nopeus ensimmäiseen viestiin on tärkeä, ensimmäinen layout saattaa olla hieman parempi vaihtoehto. Samaan aikaan näkyvien chat-ikkunoiden määrä kannattaa kuitenkin valita tarkkaan, jotta stressin taso ei nousisi liikaa.

Avainsanat asiakaspalvelu-chat, moniajo, käytettävyys, kokeellinen tutkimus

Preface

First of all, I would like to thank giosg for giving me an opportunity to write this thesis about an interesting topic. Special thanks to my advisor Kimmo Kiiski, who helped me re-organize my thoughts on the subject of this thesis many times and supported me by giving ideas, challenging my assumptions and solutions, and most of all, by encouraging me to keep going despite all the challenges I have faced along the way.

Second, I can't thank enough my advisor Jussi Jokinen for his valuable guidance and ideas on the subject and the whole thesis process. In addition, I would like to thank my supervisor Antti Oulasvirta for challenging me to come up with a meaningful subject, and for giving me guidance on the thesis process.

Even though this thesis was a huge challenge for me, I survived it better than I expected. Thank you riivatut DIPPA-mammat for being a great peer support by listening and helping me along the way.

It's hard to believe that my studies are finally over. These endless years have not been easy, but I have learned a lot about myself (and maybe something about computer science as well). I would like to thank my family for supporting me through these years. Also, thank you Team Hogwarts for being patient and for listening to my struggles through the last years. I'm all yours now.

Finally, and most importantly: 176, you have made these years special. Despite all the challenges, these years have been unforgettable and full of joy with all of you.

Espoo, 17.09.2018

Jenni Pajukoski

Contents

Abbreviations

ANOVA Analysis of Variance

CSC Customer Service Chat

HCI Human-Computer Interaction

ICC Intraclass Correlation Coefficient

ISO International Organization for Standardization

PDF Portable Document Format

UI User Interface

1 Introduction

1.1 Context

Over the last years, customer service chat has become an important part of customer service and support. Customer service chat is an internet service that allows chat agents to have real time conversations with customers visiting companies' websites (Elmorshidy 2013). Studies have shown numerous benefits for having chat as a customer service channel, such as increased conversion rates (BoldChat 2015, Kayako 2017, Kazmi et al. 2016), increased customer satisfaction (eDigitalResearch 2014, Zendesk 2015), and decreased shopping cart abandonment probability (Kang 2013, Wasserman 2001). Moreover, chat has been shown to be operationally less costly than a phone customer service (Clarkson 2010, Shae et al. 2007). Therefore, many traditional call centers have turned into "contact centers" by supplementing their phone and email support with chat (Lockwood 2017, Luo & Zhang 2013).

One of numerous chat service providers is giosg.com Ltd. Currently, giosg has 650 customers in 12 countries, with approximately 35M unique visitors in a month on customers' websites (giosg.com Ltd 2018*a*). Those visitors are potential chat end-users. Besides live chat, giosg provides other features, such as machine-learning based customer targeting, possibility to see customer's shopping cart, analytic reports, co-browsing, and many more (giosg.com Ltd 2018*b*). This thesis focuses on the live chat feature.

The basic idea of a chat service is that a chat agent can have real time conversations with one or more customers at the same time. Waiting for a response to one chat gives the agent a chance to switch context by reading and responding other chats. This is different from a traditional phone customer service, where an agent can have only one phone call at a time. A single chat is usually longer than a phone call, because both participants have to read the received messages and then write answers to them. Therefore, in order to keep the same service level than for phone calls, the ability to have multiple chats simultaneously is an important feature.

Having multiple chats simultaneously means that the chat agents are multitasking between the conversations. More and more customers consider chat as the most preferred way of contacting customer service, and therefore, the amount of chats daily is increasing. For example, evidence from giosg chat database shows that the amount of daily chats has grown 72.5% during the last two years [1]. Furthermore, Comm100 (2018) found that within their customers, the amount of monthly chats increased 80% from year 2016 to year 2017. Thus, the importance of multitasking in customer service chat is constantly growing.

Today, multitasking is a part of everyday life. Advanced technology has enabled people to talk to the phone while walking or driving a car, to cook while watching television, to check email while doing other computer tasks, and so on. The list is long. There are many reasons why people choose to multitask in the first place. According to David et al. (2013), many of them try to be efficient - they are trying

[1] Calculated from giosg chat database by taking averages from April 1st 2016 to May 31th 2016, and averages from the same accounts from April 1st 2018 to May 31th 2018

to optimize processes and to save time. Moreover, people often choose to multitask because it is more interesting and challenging than performing a single task at a time (Sanbonmatsu et al. 2013). Yet another motivation is enjoyment (David et al. 2013).

However, numerous studies have proven that multitasking decreases performance from different points-of-view, such as task completion time and amount of errors (e.g., Adler & Benbunan-Fich 2012, Borst et al. 2015, Strayer & Johnston 2001, Gillie & Broadbent 1989). In a concurrent multitasking situation, where multiple tasks are performed in parallel, limited resources of human cognition decrease performance (Wickens 1984). In addition, performing multiple tasks in sequences loads memory, and thus increases task completion times, because each task has to be retrieved from memory when resuming back to them (Altmann & Trafton 2002).

Nevertheless, studies have shown that by understanding the limitations of human cognition, technology can be designed in a way that supports multitasking situations (Jeuris & Bardram 2016, Warr et al. 2016, Yan et al. 2017). These studies revealed that user interface (UI) design decisions have an increasing impact on user's multitasking performance. The studies indicated that carefully designed user interfaces and layouts can decrease task completion times and amounts of errors.

1.2 Research problem and scope

At giosg, we want to provide a chat service that enables efficient multitasking and is pleasant to use for chat agents. To be able to improve our service further, we want to study what kind of user interface layout for chat agents supports these goals. What also affects future design decisions on chat agent user interface, is the number of simultaneous chats. Thus, we also want to study how many simultaneous conversations agents are usually having, and how different amounts of simultaneous conversations affect usability of the chat service in different layouts.

The research problem in this thesis is:

> *RP: How chat agent user interface layout affects usability in chat multitasking?*

To address the research problem, the problem is divided into three main research questions:

> *RQ1: How can usability be measured in customer service chat?*

The research problem aims to discover multitasking effects on usability. However, there are various definitions for usability in the field of human-computer interaction (HCI). Therefore, as a first step of addressing the research problem, I study literature and previous work on customer service chat and conversational multitasking in order to define how usability can be measured in customer service chat. First, I review literature in order to define usability in general. Then, I discover the most

significant variables of usability in customer service chat, and how those variables can be measured.

> *RQ2: How would layout design affect usability in customer service chat?*

To further justify the research problem, I explore multitasking literature and previous work on user interface layouts in multitasking, to discuss how layout design would affect multitasking in customer service chat. I discover two most significant layouts for customer service chat, to further investigate the effects of layout in this study.

To address the main research problem, I implement a web chat prototype and conduct an experiment with it. Based on findings for the second questions, I compare the layouts in the experiment.

> *RQ3: How does the amount of simultaneous chats affect usability in customer service chat?*

The amount of simultaneous chats affects user interface design decisions. Therefore, I discover most two significant amounts for simultaneous chats in customer service. In the experiment, I study how these amounts affect usability. Moreover, I study whether the amount of chats affects the two layouts differently.

The goal of the experiment is to study how the two chosen layouts and chat amounts affect the usability variables discovered in the first research question. Based on literature study for the second and third research questions, I formulate hypotheses for the experiment variables.

A few previous studies have investigated chat from multitasking and user interface points-of-view. However, most of them have studied only two simultaneous conversations (Dresner & Barak 2006, 2009, David et al. 2013). Wang et al. (2013) and Catanzaro et al. (2006) studied user interface effects with more than two chat windows, but both of the studies included only detecting important information or events from the chat streams. To the best of my knowledge, this is the first study where the effects of layout on chat multitasking are studied both with more than two conversations and with actually interacting with the chat windows by responding the chats. As pointed out, the amount of daily conversations in customer service chat is constantly growing. To be able to fully exploit the current resources to handle this growth, this study is an important step towards being capable to provide the best tools for chat multitasking.

1.3 Outline

The rest of this thesis is organized as follows: First, section 2 introduces customer service chat and its importance. giosg live chat service is introduced, as well as other existing providers and their differences. Second, the section studies usability in general, and in customer service chat. Most important measures for different aspects

of usability in customer service chat are discovered. The section continues by studying and introducing basic terminology and concepts of multitasking. Multitasking is related to customer service chat as conversational multitasking and previous studies on conversational multitasking are presented and discussed. In addition, theories affecting multitasking performance are introduced and applied to conversational multitasking.

Section 3 presents my hypotheses for the usability variables discovered in section 2.2. I formulate the hypotheses based on theories and previous work explored in section 2.

Section 4 presents the research method used in this thesis. The section describes the principles for constructing an experiment and for analyzing the results, along with my own construct decisions. Finally, the section describes the conducted experiment, including the design, participants, setting, and procedure.

Experiment results are presented in section 5 and discussed in section 6. Finally, section 7 concludes the thesis.

2 Background

To be able to answer the first research question, "How can usability be measured in customer service chat?", this section starts by studying literature, previous work, and data on customer service chat. First, the definition, facts, and statistics about customer service chat are introduced. In addition, current chat providers and their UI layouts are explored. Second, definitions and measures of usability are studied. Then, the most important aspects of usability in customer service chat are selected to be investigated in this study.

The latter two subsections study multitasking concept in general and in customer service chat. To be able to formulate hypotheses for the study, previous work and theories on multitasking, as well as their impact on usability, are studied.

2.1 Customer service chat

Customer service chat (CSC) allows customer service agents to have real time conversations with the customers visiting their websites. CSC is an internet service that is based on an instant messaging application. (Elmorshidy 2013) It has widely been taken as a part of companies' customer service, to supplement and replace phone calls and email (Lockwood 2017, Luo & Zhang 2013, Tezcan & Zhang 2014). Zendesk (2015) found that companies who have enabled web chat on their website are receiving less tickets through their web forms, which shows that people prefer to have real time chats while visiting companies' websites.

There are numerous advantages using chat as a customer service channel. First, CSC has been proven to be operationally less costly than phone customer service (Shae et al. 2007). In addition to savings in phone charges, costs are saved because less agents are needed in customer service (Clarkson 2010, Lockwood 2017). The reason for this is that one agent can have multiple chats simultaneously. Another benefit of CSC compared to phone customer service is faster speed to answer. Shae et al. (2007) stated that the average speed to answer to a chat is 3.5 times faster than to a phone call. CSC has also been proven to increase customer satisfaction (eDigitalResearch 2014, Zendesk 2015). In a study of eDigitalResearch (2014), live chat had the highest customer satisfaction level of all customer service channels, 73%, whereas the satisfaction level of phone customer service was only 44%.

Moreover, studies have shown numerous financial benefits for having chat as a customer service channel in online shops. Those benefits include, for example, increased conversion rates (BoldChat 2015, Kayako 2017, Kazmi et al. 2016), decreased shopping cart abandonment probability (Clarkson 2010, Kang 2013, Wasserman 2001), increased returning and re-purchasing probability (Kayako 2017), and increased amount of money spent per purchase (BoldChat 2015, Clarkson 2010). Increased conversion rates and amount of money spent per purchase can be explained by up-selling, which is possible through a web chat (TELUS International 2015). A study of Kang (2013) showed that 83% of customers wanted support during online purchase processes. They found that 48% of customers would abandon the purchase if they did not get help within a time frame, which was from immediate to 5 minutes.

Thus, offering real time help during the purchase process may decrease the shopping cart abandonment probability.

One example of chat providers is giosg.com Ltd. Founded in 2011, giosg started as a live customer service chat provider. Over the years, giosg has developed numerous other features besides the chat (giosg.com Ltd 2018*b*). Currently, giosg has 650 customer companies in 12 countries, in various industries, such as real estate, banking, insurance, contact centers, online shops, and many more. Chat agents using giosg are having an average of 15.43 chats per day (sd=38.65, $median$=6)[2].

The main functionality of giosg live chat is as follows: Chat agents can set themselves online or offline. If there are any agents online, chat can be shown for customers in the company's website. Incoming chats can be routed for different agents based on different conditions, such as customer's location on the website or customer's language. Once a customer starts a chat, anyone of the agents who the chat is routed for, can take it and start a conversation with the customer. The agents can have as many simultaneous chats as they want. All open chats are shown simultaneously in separate chat windows for the agents, but the chat windows can also be minimized to the left side of the browser window.

Besides giosg, there are several other CSC providers in the market (e.g., Comm100 Network Corporation 2018, Intercom 2018, LiveChat Inc 2018, LivePerson Inc 2018, Vergic 2018, Zendesk 2018). Those providers differ from each other by their user interfaces and available features. Browsing through different chat providers revealed that there are clearly two different types of layouts for chat agent user interface: In the first one, all of the simultaneous chats are shown at the same time for the agent in separate chat windows (e.g., giosg.com Ltd 2018*c*, Vergic 2018) (see Figure 1). In the second one, only one conversation is shown at a time, and it can be changed from a list of all simultaneous conversations (e.g., Intercom 2018, LiveChat Inc 2018, Zendesk 2018) (see Figure 2). Even if the layouts and features differ between chat providers, they all provide ability to have multiple chats simultaneously.

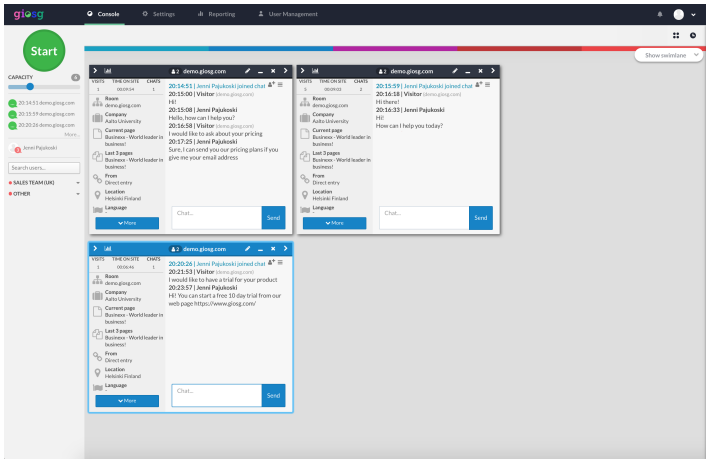


Figure 1: Chat agent user interface layout in giosg live chat (giosg.com Ltd 2018*c*).

[2]Calculated from giosg chat database as daily averages between April 1st 2018 and May 31th 2018

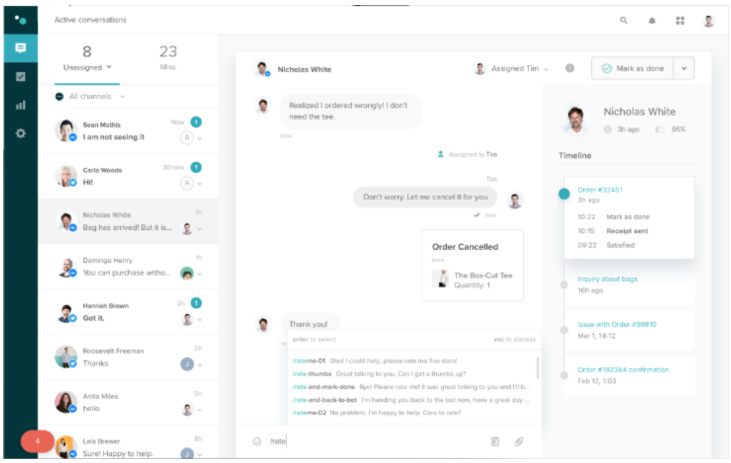


Figure 2: Chat agent user interface layout in Zendesk chat (Zendesk 2018).

The ability to have multiple simultaneous chats is a significant feature in customer service, because a chat with a customer usually takes more time than a phone call (Lockwood 2017, Shae et al. 2007). The reason for this is that a chat requires both the agent and the customer first to read the received message and then to write an answer. Therefore, in order to serve the same number of customers in the same time as with phone customer service, agents have to be able to attend multiple chats simultaneously.

Data from giosg chat database showed that on average, the users have 2.30 (sd=2.20, $median$=1) simultaneous conversations [3]. Moreover, Comm100 (2018) benchmark data showed that agents were having an average of three simultaneous conversations. A study of Shae et al. (2007) showed that a chat agent could handle three simultaneous chats without a decrease in performance. However, they did not investigate performance with more than three conversations, and therefore I want to study also four simultaneous conversations in this thesis.

Having multiple simultaneous chats means that the agents are multitasking. Data from giosg database show that the amount of daily chats has grown 72.5% during the last two years[4]. The growing amount of chats daily is increasing the importance of the ability to multitask efficiently.

The layouts introduced above are quite similar, if the agent is having only one conversation at a time. However, having multiple simultaneous chats may have different effects on usability between the two layouts. In the first layout, the agent sees all conversations at the same time, which may interfere concentration on one conversation. On the other hand, in the second layout, extra click is needed if the agent wants to see some other conversation. The following subsections introduce theoretical backgrounds for multitasking, and how those theories may affect usability in customer service chat. After that, section 3 presents my hypotheses on how the theories may affect usability in the two layouts, with different amounts of simultaneous

[3]Calculated from giosg chat database between April 1st 2018 and May 31th 2018

[4]Calculated from giosg chat database by taking averages from April 1st 2016 to May 31th 2016, and averages from the same accounts from April 1st 2018 to May 31th 2018

conversations.

2.2 Usability in customer service chat

To address the research problem of this thesis, I first need to define what is *usability* and how it can be measured. According to ISO (International Organization for Standardization) 9241 standard (ISO 2010), usability is defined as "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Other suggested definitions for usability are "the capability to be used by humans easily and effectively" (Shackel 1991) and "quality in use" (Bevan 1995).

Hornbæk (2006) stated that besides defining usability, another challenge is how to measure it. According to him, it cannot be directly measured. Instead, it has to be operationalized into aspects that can be measured. In his study, Hornbæk (2006) reviewed 180 studies from HCI literature, in order to understand the current practice in measuring usability. He grouped the measures according to the ISO 9241 (ISO 2010), into *effectiveness*, *efficiency* and *satisfaction*. These are the aspects of usability that I am using in this study as well.

ISO 9241 (ISO 2010) defines effectiveness as "accuracy and completeness with which users achieve specified goals". Hornbæk (2006) found that the most common measures of effectiveness were *binary task completion*, *accuracy*, *recall*, *completeness* and *quality of outcome*. Binary task completion was measured as if the user has completed the tasks or not, and accuracy as the amount of errors during tasks. Recall was measured as the amount of information user was able to recall after interaction, and completeness as the extent to which the tasks were solved. Finally, quality of outcome was measured as a quality of work or learning, for example.

In ISO 9241 (ISO 2010), efficiency is defined as "resources expended in relation to the accuracy and completeness with which users achieve goals". According to the literature review by Hornbæk (2006), there were six common measures for efficiency: *Time* was used to measure task completion times or time to complete parts of the tasks. *Input rate* was usually measured as text entry speed. *Mental effort* referred to the mental resources used while completing tasks, and it was often measured with NASA Task Load Index (Hart & Staveland 1988) questionnaire. *Usage patterns* were used to measure how the interface was used, and *learning measures* as the effects of learning during usage of the interface, in terms of changes in task completion time, for example. *Communication effort* measured the resources expended in communication during tasks.

Finally, in ISO 9241 (ISO 2010), satisfaction is defined as "freedom from discomfort and positive attitudes towards the use of the product". According to Hornbæk (2006), 22% of the studies used *preference* as a satisfaction measure. Preference indicated which user interface the users preferred. *Ease-of-use* was a broad measure for overall satisfaction or *attitudes towards the interface*. Most common measures for attitudes were *liking*, *fun*, *annoyance*, and *control*. Yet another measure for satisfaction was *attitudes towards other persons*, which measured attributes like feeling of *presence*, *trust*, *common ground*, and *ease of communication*. Typically, all

measures of satisfaction were measured with *Likert scale questionnaires* (Likert 1932). Likert scale questionnaires contain statements about the interface, for example, "I think the interface was fun to use". Each of the statements is answered with a scale, usually from 1 to 5 or from 1 to 7, 1 meaning "strongly disagree" and 5 (or 7) meaning "strongly agree", for example. According to Hornbæk (2006), 7% of the studies in the review used standard questionnaires to measure satisfaction.

Usability in customer service chat can be measured in many different ways. Many studies have investigated CSC in terms of efficiency. The most common metric seems to be response time (McLean & Osei-Frimpong 2017, Shae et al. 2007, TELUS International 2015). Moreover, length of chat session (Kang et al. 2015, Kulbytė 2018, Shae et al. 2007) and time between interaction (Shae et al. 2007) have been used to measure efficiency of CSC. However, even though response time is stated to be important, studies have shown that customers are actually willing to wait for high-quality support (Comm100 2018, Kayako 2017, Kazmi et al. 2016). Moreover, an important measure is first call resolution (i.e., whether the issue of the customer is solved within the first contact) (Kang et al. 2015, Kulbytė 2018, Shae et al. 2007). These two measures refer to effectiveness. Another suggested measure of CSC is customer satisfaction (Shae et al. 2007, Steele 2017, TELUS International 2015), but this thesis focuses on the chat agent side, end-user side is left out of scope.

All of the studies listed above suggested measuring customer service chat only from the effectiveness and efficiency aspects of usability. It is reasonable, because those measures are related to cost savings and customer satisfaction. However, as stated earlier, at giosg we want to provide a service that is also pleasant to use for the agents. Therefore, I want to measure also satisfaction in this study.

To study usability in this thesis, I decided to measure the following factors: First, efficiency is measured by *response time* and *chat duration*. I further operationalize response time to *first response time* and *question response time*, to see whether there are any differences in responding behavior in different situations between the layouts. Second, effectiveness is measured as *accuracy*. Finally, satisfaction is measured by subjective perception of *efficiency*, *stress*, *control*, *frustration*, and *retention*, as well as *preference* between layouts.

2.3 Multitasking in chat environment

Multitasking means ability to perform two or more tasks at a time, either in parallel or with frequent switches. David et al. (2013) stated that people have various motivations for multitasking. According to him, a common one is efficiency: people try to optimize processes or time to complete tasks. In addition, David et al. (2013) suggested that another motivation is enjoyment. In fact, Sanbonmatsu et al. (2013) proposed that people often choose to multitask because of enjoyment, even though their performance is negatively affected. For example, I am listening to music while writing this thesis, even though it distracts my concentration a little. Sanbonmatsu et al. (2013) stated that people are able to achieve more goals and experience more activities through multitasking. They argued that people often choose to multitask because it is more interesting and challenging than performing a single task at a

time. A study of González & Mark (2004) revealed that information workers spend average of three minutes on one task before switching to another.

In this thesis, the interest is on customer service chat multitasking, which can be called *conversational multitasking*. Conversational multitasking can be defined as participating in two or more conversations at the same time, either passively or actively (Dresner & Barak 2009). In CSC, the motivation for multitasking is the ability to service multiple customers simultaneously, which increases the total amount of customers being served with the same number of agents. The reason why multiple customers can be served simultaneously, is that there is always some idle time in chats, during which the customers are reading the responses and preparing their answers. In fact, data from giosg chat database shows that an average duration of a chat conversation (time from first to last message) is 11 min 54.57 s (sd=25 min 40.41 s, $median$=5 min 48.80 s), from which an average of 4 min 46.61 s (sd=4 min 34.16 s, $median$=3 min 37.91 s) is active time (any time between messages that exceeded 1 minute was considered idle time)[5]. This means that the agents have time to handle other customers at the same time.

However, due to cognitive limitations, multitasking often decreases performance. The effects of multitasking on performance have been studied in various different contexts, such as multitasking while driving (Salvucci 2001, Salvucci & Taatgen 2008, Strayer & Johnston 2001), multitasking in human-computer tasks (Adamczyk & Bailey 2004, Adler & Benbunan-Fich 2012, Bailey & Iqbal 2008, Bannister & Remenyi 2009), and media multitasking among students (Bowman et al. 2010, Ellis et al. 2010, Hembrooke & Gay 2003, Lee et al. 2012) and information workers (Aral et al. 2006, Cutrell et al. 2000, Czerwinski et al. 2000). Many of those studies have shown an inverted U-relationship between multitasking and productivity (Adler & Benbunan-Fich 2012, Aral et al. 2006, Bannister & Remenyi 2009). This means that multitasking can be productive to some extent, but the productivity decreases when a certain threshold is exceeded. An inverted U-relationship has also been found between multitasking and motivation, ability, and opportunity (Bardhi et al. 2010). However, increased productivity caused by multitasking has been proven to decrease accuracy (Adler & Benbunan-Fich 2012). Multitasking has also been shown to reduce retention and topic interest (Dindar & Akbulut 2016).

The effects of conversational multitasking on performance have been studied previously in different contexts. Dresner & Barak (2006) studied effects of different user interfaces in multitasking between two conversations. They compared three conditions: two conversations in two distinct windows, two conversations intertwined in one window, and two conversations intertwined in one window, but distinguished through color. They tested how well the participants could follow the conversations, and tested the number of correct answers (i.e., accuracy) with multiple-choice questionnaires after the conversations. Their results showed that separating conversations into two distinct windows helped people the most to follow the conversations, and that within one window, color separation was more effective than no separation. In another study, Dresner & Barak (2009) again studied the effects of user interface

[5]Calculated from giosg chat database as averages between April 1st 2018 and May 31th 2018

in two conversations multitasking situation. This time, they measured the number of correct answers in two experiments: In the first experiment, they tested the conversations in two separate windows, with two conditions - approximate windows and distant windows. In the second experiment, they had four conditions - two conversations in one window and in two windows, and both of them with long and short windows. The results showed that the distance between the two windows made no difference in accuracy. Moreover, separation in two windows resulted higher accuracy in both window sizes, and the difference between results was higher with long windows than with short windows.

The effects of user interface with more than two simultaneous conversations have also been investigated in a few studies. Wang et al. (2013) compared two user interfaces in a situation, where critical words had to be written down from ten chat streams in ten distinct chat windows. In addition, facts about the contexts had to be remembered after the experiment. The first UI displayed chats in square windows, where seven lines of text were displayed at a time. The second UI displayed only one line of text, that was flowing continuously from right to left. They measured the number of detected words and the number of reported facts, and the results showed that the first UI worked better. In a post-experimental survey, the first UI was preferred, and the participants reported that in the first UI the new messages in chat windows were easy to detect by looking for movement. The researchers did not test spatial memory in the experiment, but they assumed that it might have had a role in the results.

Catanzaro et al. (2006) tested detecting critical events in multiple chats with similar user interface layouts that I found popular among CSC providers: *tiled*, where all the chat windows where visible simultaneously, and *tabbed*, where only one window was visible at a time. Both layouts were tested with text highlighting and with no highlighting, and with fast and slow message rates. They measured efficiency as the time to detect critical events and the subjective mental workload, and effectiveness as the percentage of detected events and the percentage of false-alarms. The results indicated that there was no significant difference in detecting events between the layouts. However, highlighting had a slight disadvantage in the tabbed layout, and a slight advantage in the tiled layout, as measured by response time. In addition, comments by participants revealed that in the tabbed layout, it was difficult to determine where they had left off, and that new messages in the tiled layout were easy to detect through motion.

To be able to understand the effects in the studies and to design user interfaces that support multitasking, the behavior and cognitive theories in multitasking situations have to be understood. There are different types of multitasking behavior. Salvucci et al. (2009) presented a theory of *multitasking continuum*. In their theory, they characterized multitasking behavior as "the time spent on one task before switching to another". Figure 3 shows the time span for switching frequency, and typical tasks for those frequencies.

The type of multitasking where switching happens almost every second or even more frequently (left side of the span in Figure 3), is called *concurrent multitasking* (Salvucci et al. 2009). Practically, both tasks are performed at the same time.

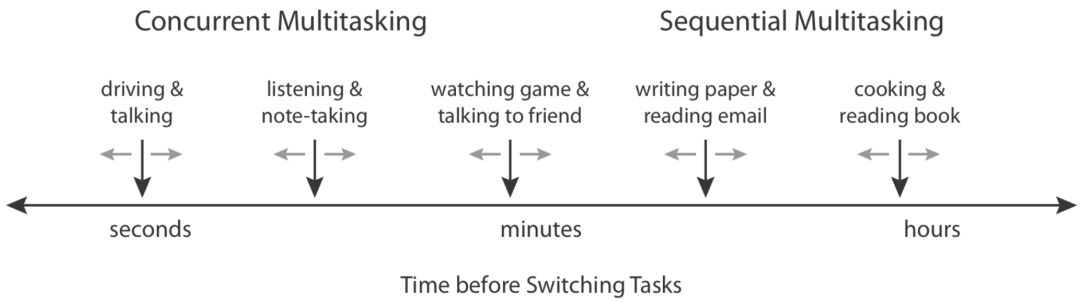


Figure 3: Multitasking continuum, where multitasking behavior is presented as a time span for switching frequency (Salvucci et al. 2009).

Typical examples of concurrent multitasking are driving and talking to a phone at the same time, or listening to a lecture and taking notes. There is one primary task (e.g., driving) that is interrupted by a secondary task (e.g., talking to the phone). Both tasks can be performed in parallel, but the performance is limited by different resources (Salvucci et al. 2009).

Behavior on the right side of the span in Figure 3 is called *sequential multitasking* (Salvucci et al. 2009). In sequential multitasking, switching happens less frequently, from minutes to hours even. According to Jeuris & Bardram (2016), unlike in concurrent multitasking, there are no secondary tasks in sequential multitasking. Instead, multiple primary tasks are interleaved.

In sequential multitasking, switching between tasks is caused by interruptions. Primary task can be interrupted either internally or externally (Katidioti et al. 2016). According to Katidioti et al. (2016), *internal interruption* is self-initiated - the user voluntarily switches between tasks, whereas *external interruption* is caused by another task. Typical example of an external interruption is a system notification or a phone call.

Roughly speaking, customer service chat with multiple simultaneous conversations can be considered as sequential multitasking. Each chat is as a primary task and only one of them can be concentrated on at a time. Switching between chats is done either as a result of external interruptions (notifications) or internal (self-initiated) interruptions. External interruptions in chat are the notifications for new conversations or new messages. An example of internal interruption is that when a new message arrives, the agent asks the customer to wait a moment, and then later decides to answer the chat without a new notification.

Numerous studies have shown that interruptions are disruptive in terms of usability. First, interruptions affect efficiency by increasing task completion time, because it takes time to recover from an interruption (Borst et al. 2015, Cutrell et al. 2000, Monk et al. 2008). Longer execution times are caused by *interruption lag* and *resumption lag* (Salvucci et al. 2009). Interruption lag means the time from the interruption to the start of the interrupting task, whereas resumption lag is the time from the end of the interrupting task to the restart of the interrupted task. Second, from the effectiveness perspective, the probability of making errors in the interrupted

task is increased after interruption (Li et al. 2008). One cause for errors is that after an interruption, a wrong task is retrieved from memory (Altmann & Trafton 2002).

Even though external interruptions have been criticized as being disruptive, Katidioti et al. (2016) showed that self-initiated interruptions can be even more costly in terms of time. The resumption lag did not differ between interruptions in their study, so they reasoned that the additional time should be caused by decision-making. Thus, external interruptions can be considered as a task management tool, which was stated also by Paul et al. (2015). However, to be a good tool for task management, notifications have to be designed carefully. The content, length, type, salience and timing all have influence on the level of the disruptive effect of the notification (Adamczyk & Bailey 2004, Gillie & Broadbent 1989). Paul et al. (2015) also suggested that switching to the interrupting tasks should be provided through the notification in order to minimize disruption on the primary task.

In chat environment, external notifications are important. Without any notifications, chat agents would constantly have to browse through the open chats in order to notice new messages. However, internal notifications are also an essential part of conversational multitasking. There are situations where agents cannot react to the new messages immediately, which means that they need to remember and decide to do that later.

Even though sequential multitasking with notification interruptions may sound reasonable characterization for conversational multitasking, the reality is more complex. To some extent, chat agents are usually able to write a message, while simultaneously reading a new, incoming message. If multitasking is only considered to be divided into concurrent and sequential multitasking, this behavior would be concurrent multitasking, and CSC multitasking would be a mixture of both of them.

A more reasonable way of modeling chat multitasking would be to think about the resources that are needed for it. First of all, *long term memory* is needed to remember the content of each conversation. When chat agents are switching between multiple conversations, they need to remember what each conversation is about in order to continue it. Otherwise, they need to read the conversation again when switching back from other conversations. Furthermore, if agents decide to postpone answering some messages, while answering some other messages, memory is needed to remember that later. Second, *attention* and *perception* are needed to notice new messages. Attention may be visual or aural, depending on the type of the notification. If no notifications are shown, visual attention may still be able to discover a new message through motion. While a new message draws attention away from a conversation, performance on that conversation is disturbed. This is because certain resources cannot be divided, which is the third aspect of modeling chat multitasking. In order to explain multitasking effects on customer service chat multitasking, I study theories behind memory, visual attention and available resources in the next subsection.

2.4 Applied multitasking theories

Increasing and decreasing performance effects of multitasking can be explained by various theories. This subsection introduces the main theories that explain concurrent

and sequential multitasking, and how these theories can be applied to conversational multitasking. With these theories, I try to explain how multitasking affects usability in chat environment.

2.4.1 Multiple resources

Most of the people have probably noticed that in dual or multiple task situations, some task combinations distract each other more than some other combinations. To explain that, Wickens (1984) proposed a theory of multiple resources. According to the theory, people have a limited set of resources available for mental processes. Multiple resource theory suggests that those resources are divided into four dimensions: *processing stages*, *perceptual modalities*, *visual channels*, and *processing codes*. Each dimension has two discrete levels. Figure 4 show the four dimensions and their levels. The model proposes that if two tasks are using resources from a same level, their time-sharing performance is limited. (Wickens 1984)

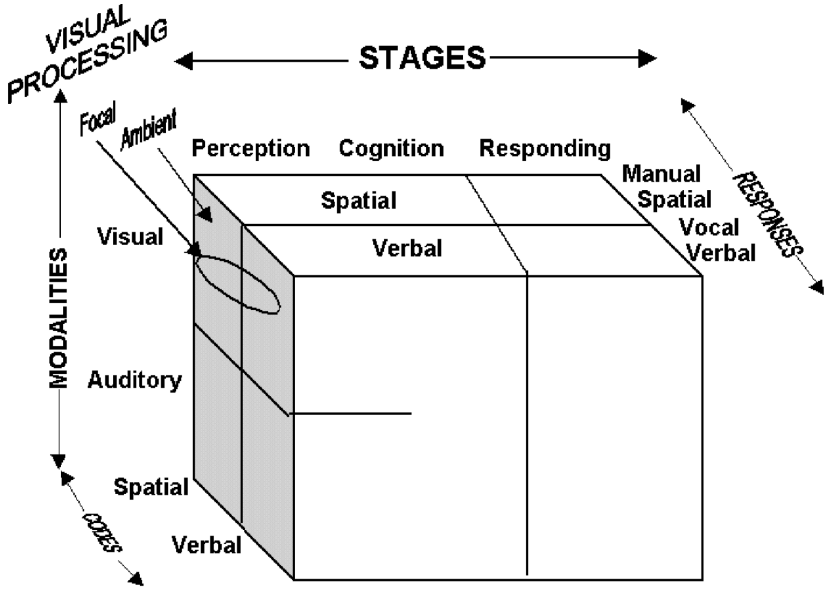


Figure 4: Four dimensions of multiple resources model. The fourth dimension is nested within visual resources in this three-dimensional representation. (Wickens 2002)

The two levels of processing stages are perception and cognition, and responding. Perceptual modalities are divided into visual and auditory levels. Visual modalities are further divided into two visual channels: focal and ambient. The fourth dimension, processing codes, are divided into spatial and verbal levels. (Wickens 1984)

Multiple resources model can be used to explain some time-sharing performance issues in chat multitasking. First of all, two chats cannot be read at the same time, because they both would need the focal visual channel. However, while reading one chat, it is possible to notice new messages in another chat, because it requires the ambient visual channel. Second, even though people may be able to write text

without looking, reading other text at the same time interferes the writing. This is because both of those processes use resources from verbal processing codes. Thus, according to the model, multitasking affects efficiency in terms of increased time to respond the conversations.

2.4.2 ACT-R and threaded cognition

ACT-R (Adaptive Control of Thought-Rational) (Anderson 1993) is a theory and a computational model of human cognition. It can be used to simulate cognitive tasks and predict their outcomes. According to the theory, cognition is a set of modules that are independent but interact with each other. Central cognitive modules include *declarative memory module*, *goal module*, *problem representation module* and *procedural module*. The declarative memory module stores factual knowledge, including task instructions. The goal module stores the current goal and progress and state information. The problem representation module maintains partial representations of the task. Finally, the procedural module integrates information from all the modules and maps these into actions that are then executed by other modules. Other modules include perceptual and motor modules that deal with vision, audition, manual control, and speech, for example. (Salvucci & Taatgen 2008)

The idea of the ACT-R theory is that the modules can operate in parallel, but each module can hold only one task at a time. Thus, bottlenecks on performance appear if two tasks are trying to use the same module at the same time. (Salvucci & Taatgen 2008)

Threaded cognition (Salvucci & Taatgen 2008) was built on top of ACT-R. Whereas ACT-R assumed that a single goal could be in the goal-buffer at a time, Salvucci & Taatgen (2008) suggested that multiple active goals could be added and maintained in the goal-buffer. Thus, threaded cognition was provided to model and predict performance during concurrent multitasking.

The main idea of the threaded cognition model is that the set of active goals are executed as threads in available modules. It suggests that a procedural module coordinates the threads, combines their outcomes from other modules, and initiates new goals in those modules. (Salvucci & Taatgen 2008)

Because the models are developed for concurrent multitasking, as an example of chat multitasking, we can consider the following situation: The agent is first focusing on one conversation, and then a new message is received to a second conversation, while the agent is writing an answer for the first conversation. The agent reads the new message in the second chat and tries to continue writing in the first conversation at the same time.

ACT-R and threaded cognition models can be applied into the example situation as follows: First, the visual module detects a new message and the procedural module instructs to attend to the message. Then, the visual module encodes the message, at the same time as the procedural module retrieves syntax for the words from the declarative memory. The procedural module attaches the syntax and guides the problem representation module and/or the declarative memory module to form an answer for the message. At the same time as the answer is forming, or after that, the

procedural memory guides to initiate writing, which is then done by a manual control module. If a new message is received in a second conversation during writing, the visual module again detects that and the procedural module directs attention to the new message. The new message is added to the set of active goals in the goal module. Like when receiving the message in the first conversation, the procedural, visual, and declarative modules are needed to encode and understand the message in the second conversation. Thus, because the agent simultaneously continues writing in the first conversation, both tasks have to take turns waiting for the procedural, declarative memory, and problem representation modules. From usability point-of-view, this behavior affects efficiency, by increasing response times.

2.4.3 Memory for goals

While having multiple simultaneous chats, a new message in one chat may interrupt conversation in other chat. The focus may be shifted to the other conversation for a while. When resuming back to interrupted chat, it may take a while to be able to continue the conversation. The reason for this is that the conversation needs to be retrieved from memory. Altmann & Trafton (2002) proposed a theory of memory for goals, which can be used to explain how memory works in multitasking situations.

In their theory, Altmann & Trafton (2002) defined goal as "a mental representation of an intention to accomplish a task, achieve some specific state of the world, or take some mental or physical action". They proposed that to direct behavior, a goal must be the most active goal in memory.

According to the model, there are three concepts that affect the memory behavior. First, *interference level* is the expected activation of the most active distractor. Interference level determines which goal is retrieved from memory. If the target goal is above the interference level, it is retrieved as the most active goal. Otherwise, some of the distractors is the most active goal and it is retrieved from memory instead of the target. Second, *strengthening constraint* predicts the time to encode a new goal. It defines activation as a retrieval frequency history of the goal. This means that if the activation of a goal is very high, it distracts other goals from being the most active ones. Last, *priming constraint* suggests that two components are responsible for old goals being activated again. Those components are the history of the goal, i.e., the retrieval frequency, and cues in the mental or environmental context. If a cue for the goal is presented, priming reactivates an old goal and that goal may direct behavior again. (Altmann & Trafton 2002)

In chat multitasking, every new chat can be considered as a new goal. A goal could be to help to find a jacket for the customer. Each chat can include sub goals, such as answering questions about the available jackets. Once a chat is started, its activation is strengthened above the interference level of other chats. Old chats may be reactivated in memory with notifications as cues. A priming cue may also be visibility of a chat. Altmann & Trafton (2002) proposed that sometimes long-term knowledge about a task can act as a cue to direct behavior. While chatting with customers, agents have a strong knowledge about that they need to serve the customers in the open chat windows. Thus, notification is not always needed to

retrieve a specific chat goal from memory.

According to the model, retrieving old goals from memory takes time. In a chat multitasking situation, this means that efficiency is affected by increasing time to respond the conversations. In addition, effectiveness is affected by increased probability for errors, because multiple chats interfere each other. After an interruption, a wrong chat may be retrieved from memory and cause errors.

2.4.4 Visual attention

Visual attention plays an important role in conversational multitasking. Other chats and notifications should not be too disruptive for the chat that is currently concentrated on. However, without any notifications or cues for new messages, chat agents should actively browse the chats in order to notice them. In fact, Katidioti et al. (2016) proposed that external interruptions are less disruptive than self-interruptions. They did not find any difference in task resumption time, hence they reasoned that the longer task completion time must be caused by decision-making. Thus, it is important that the layout supports visual attention in a reasonably disruptive manner.

Visual attention can be divided into *bottom-up attention* and *top-down attention* (Itti & Koch 2001). Bottom-up attention is based on visual saliency. According to Wolfe & Horowitz (2004), attributes that most probably guide bottom-up attention are color, motion, orientation and size, whereas color change, intersection and semantic category, for example, are not guiding attributes. Itti & Koch (2001) stated that saliency seems to be derived by feature contrast. Therefore, color contrast, motion, orientation contrast and size contrast should be used in order to create salient notifications.

However, it matters how those attributes are used. For example, if there are already many colors in a user interface, adding some other color in order to drive attention probably does not work. The reason is that the UI is *cluttered* (Rosenholtz et al. 2007). The added color does not pop out from a cluttered UI as probably as from a non-cluttered UI. The same effect applies to other attributes as well.

The other part of visual attention, top-down attention, is driven by task-specific cues and previous knowledge (Itti & Koch 2001). In a chat context it means that the agents are able to find the new messages from other chats quickly because they know where to look for. In addition, if the color of new message notifications is always red, the agents know to look for red in the UI in a top-down manner.

Visual attention can have a huge impact on usability, in terms of efficiency and effectiveness. The reason is that disruptive notifications from other chats guide bottom-up visual attention away from the current chat. From the other theories and models introduced above, we can conclude that these interruptions affect efficiency by increasing response times and effectiveness by increasing the possibility for errors. Moreover, without any cues for visual attention, response time can also be increased because of decision-making.

3 Hypotheses

Hypotheses are formulated to predict effects of the independent variable manipulations on the dependent variables. Good hypotheses are concise, justified and testable. Justification should be based on previous work, theories, and models. Benefits of hypotheses are that they clarify the focus of the research question, help summarize previous work, and to report the experiment. However, hypotheses can sometimes be formulated to be disconfirmed. (Hornbæk et al. 2013)

This section presents formulation and justification of my hypotheses on this study. In the previous section, I found two chat UI layouts that I investigate further in this study. In the first layout, all of the chat windows were visible at the same time. In the second layout, only one chat window is visible at a time, and the other conversations are listed on the left side of the browser window. In this thesis, I call these layouts *windowed layout* and *tabbed layout*, respectively.

Based on literature review, Wickens et al. (2015) studied task switching behavior. They proposed a model for how task difficulty, salience, priority and interest affected switching and avoiding switching behavior. The model suggests that people are more likely to switch to easier tasks than to difficult tasks. Moreover, according to the model, saliency of alternative tasks makes people switch to them more likely.

As mentioned earlier, Catanzaro et al. (2006) tested similar, tiled and tabbed layouts for multiple chat windows. They found that it might be easier to detect critical events form the tiled layout, because of the saliency of new messages. In my study, the windowed layout corresponds to the tiled layout.

Switching between chats in the tabbed layout requires an extra click, and thus is more difficult than in the windowed layout. Furthermore, I assume that in the windowed layout, new messages are more salient and thus easier to detect than in the tabbed layout. Therefore, I assume that more interruptions and more switching between chats happen when using the windowed layout.

Moreover, while writing a message to a chat, in the windowed layout the user sees other conversations, and may try to read them at the same time. According to ACT-R and threaded cognition theories, this may increase response times in the windowed layout.

In the literature review, I found two different amounts of simultaneous chats that I investigate with both of the layouts. The amounts are three and four. I assume that there are more interruptions and that the users have to switch more between chats with four chats. In addition, Altmann & Trafton (2002) suggested that in a multitasking situation, a wrong goal may be retrieved from memory after an interruption, if the activation of the target goal is below interference level. With four simultaneous chats, there are even more distracting goals. Thus, the probability to retrieve a wrong goal is higher with four chats than with three chats.

Based on these assumptions, my hypotheses for the usability measures in this study are the following:

H1: First response time is faster in the windowed layout.

According to Duggan et al. (2013), users are interleaving tasks in order to maximize the marginal rate of return. Because response time is an important measure in chat, I assume that users try to minimize that. Responding the first message is usually a quick and easy task, if it is a greeting. Thus, I predict that users are likely to switch to that task, while doing other tasks. However, because switching is easier and faster in the windowed layout, my hypothesis is that first response time is faster in the windowed layout than in the tabbed layout.

> *H2: There is no significant difference in question response time between the windowed layout and the tabbed layout.*

As stated above, switching time is slower in the tabbed layout. From that point-of-view, I would assume that responding to questions would be slower in the tabbed layout. However, I suggest that there are also other factors affecting the response time. First, as assumed above, more switching between chats is done in the windowed layout. Therefore, responding to a single question could take more time in the windowed layout. Moreover, as other chats are more salient in the windowed layout, they probably cause more disruptive interruptions that draw participants' visual attention and make participants more likely to read the other chats. Thus, from that point-of-view, answering the questions might take more time in the windowed layout. However, Salvucci et al. (2009) stated that the visibility of the interrupted task rehearses it in memory, and therefore, resuming it does not take as much time as to a task that has not been visible during interruption. According to that, if an interruption happens in the tabbed layout, it takes more time to resume to the interrupted task than in the windowed layout. Because of all these factors, I predict that there is no significant difference in question response time between the windowed layout and the tabbed layout.

> *H3: Accuracy is higher in the tabbed layout.*

As stated in memory for goals theory, errors are more likely after interruptions. I assumed that more switching is done in the windowed layout, and therefore I predict that more errors are done in it. In addition, interruption lag is higher with the tabbed layout, as it takes time to switch to a conversation. Many studies have shown that a longer interruption lag decreases the probability of errors after switching. Thus, I expect that the accuracy is higher in the tabbed layout.

> *H4: There is no significant difference in chat duration between the windowed layout and the tabbed layout.*

For the same reasons as for H2, I suggest that there is no significant difference between chat duration between the windowed layout and the tabbed layout. Even though the first response time would be faster in the windowed layout, I predict that the higher switching frequency in the windowed layout makes the chat duration similar for the windowed layout and the tabbed layout.

H5: There is no significant difference in first response time between three and four simultaneous chats.

I predicted that the first response time is faster in the windowed layout than in the tabbed layout. I assume that the same reasons affect the first response time no matter how many simultaneous chats there are. Therefore, I predict that there is no difference in first response time between three and four simultaneous chats.

H6: Question response time is slower with four simultaneous chats.

Because I assume that the first response time would be the same for three and four chats, I therefore predict that there will be more switching and interaction with four chats than with three chats. Moreover, there are more chats to react and response anyway. I suggest that this makes the question response time slower in four simultaneous chats.

H7: Accuracy is higher with three simultaneous chats.

Because I assume that four chats will cause more switching and interaction than three chats, I predict that accuracy will be higher with three simultaneous chats. I assume this, because like stated earlier, errors are more likely after interruptions. Moreover, as stated above, more simultaneous chats give more possibilities to retrieve a wrong chat from memory and cause errors.

H8: Chat duration is higher with four simultaneous chats.

Based on the assumption of slower question response time in four simultaneous chats, I predict that the chat duration is also higher with four simultaneous chats.

H9: the tabbed layout causes less stress than the windowed layout.

In their computer desktop workspace study, Jeuris & Bardram (2016) found that the cognitive load was higher in Windows 7 environment, where all tasks are in one workspace, than in dedicated workspaces. They suggested that "Since information overload can cause stress, the mere visibility of previously suspended tasks could potentially increase perceived time pressure". Therefore, I suggest that the tabbed layout causes less stress for the participants.

To study these hypotheses, I need a method that can be used to measure response times and accuracy. Therefore, an interview or a questionnaire is not sufficient in this study. I need a method, where people are actually using a chat, to be able to measure the variables. Implementing the layouts to an existing chat software and collecting data from it would be one solution. However, it would take time to collect reliable data, because conditions such as the amount of incoming chats, time between incoming messages, and the difficulty of the chats, would have such a large variation. To get comparable results in a short time, those conditions should be controlled.

4 Method

Recall the goal of this thesis, which is to study how different layouts in customer service chat user interface affect usability in chat multitasking. As a result of the literature review, usability was operationalized as effectiveness, efficiency and satisfaction. In customer service chat context, those variables were further operationalized as follows: Effectiveness is measured as accuracy, efficiency as response time and chat duration, and satisfaction as perception of efficiency, stress, control, frustration, and retention, as well as preference between layouts. The literature review revealed that there are clearly two popular types of chat agent user interface layouts in existing CSC solutions. In addition, I found that three and four could be the most significant amounts of simultaneous chats to study the effects on usability. To study exactly these conditions with the usability measures listed above, I had to choose the method in a way that I could control the conditions, as well as conditions such as time between incoming messages and difficulty of the chats. Thus, I chose an *experiment* as my research method.

Cook et al. (2002) describes experiment as "a study in which an intervention is deliberately introduced to observe its effects" (p. 12). Hornbæk et al. (2013) characterized the concept of intervention as "a level of an independent variable, or as a treatment, or as a condition". In an experiment, a research problem is studied by testing different conditions, and sets of those conditions are called *independent variables* (Purchase 2012, p. 8). Experiments are conducted with the conditions, and the effects of the conditions are measured as *dependent variables* (Hornbæk et al. 2013). Based on previous studies and theories, hypotheses are formulated to predict variations in the dependent variables and to explain the phenomena. After conducting the experiment, hypotheses are tested against the results.

Using experiments in research has many benefits. First, experiments and their results are easy to replicate, which is an important factor when assessing validity of the research. In a laboratory experiment, it is possible to minimize the biases caused by the environment and possible random effects outside the studied phenomenon. (Gergle & Tan 2014) As a time-saving method, experiments can be used for investigating technology without deploying it (Hornbæk et al. 2013). In addition, experiments provide quantitative data that can be analyzed using inferential statistics (Gergle & Tan 2014).

A major challenge in experiments is their generalization to real word, i.e., to other people, environment, and technologies (Gergle & Tan 2014). Experiments should be designed in a way that the tasks would be realistic, but at the same time the dependent variables should be possible to be reliably measured from the tasks. Another challenge for laboratory experiments is the motivation of the participants (Hornbæk et al. 2013). If the participants are not motivated doing the tasks, it may affect the results.

This section presents the principles of experiment construction, along with my construct decisions. First, validity of a research, and how it can be assured when designing an experiment, are discussed. Second, selecting and measuring independent and dependent variables are described. Third, the experiment is described in detail,

including the design, participants, setting, and procedure. Finally, the basics of statistical analysis are described, as well as my selections for analysis methods.

4.1 Validity

When designing an experiment, it is important to assess the validity of the set-up. According to Cook et al. (2002), there are four types of validity: *statistical conclusion validity*, *internal validity*, *construction validity* and *external validity* (p. 38).

Statistical conclusion validity is related to the validity of inferences between two variables. The experimenter can incorrectly conclude a covariance between the variables, or that the covariance is considerably higher or lower than it really is. (Cook et al. 2002, p. 42) Examples of threats to statistical conclusion validity are low statistical power, violated assumptions of statistical tests, and unreliability of measures (Cook et al. 2002, p. 45). A potential threat to statistical conclusion validity in my study is that the number of participants may be too small.

Internal validity concerns whether the covariance between two variables is actually caused by the experiment manipulations instead of something else (Cook et al. 2002, p. 53). Typical threads to internal validity are selection, history, maturation, and testing (Cook et al. 2002, p. 55). In my experiment, I make sure that the participants do not get tired by allowing them to have breaks between experiment blocks. In addition, the effects of learning are minimized by assigning conditions for participants in varied orders. One potential threat to internal validity in my study is that the participants may not perform the tasks seriously.

Construct validity is about the extent to which the collected measures characterize the study operations (Cook et al. 2002, p. 65). Threats to construct validity are, for example, inadequate explication of constructs, mono-operation bias, mono-method bias, and experimenter expectancies (Cook et al. 2002, p. 73). For example, if I would measure only satisfaction, it would cause mono-operation bias, and the results would not tell the whole truth about usability. Moreover, the only method should not be asking participants subjectively about the usability, because they might feel more efficient, for example, than they really are. Thus, other methods are needed as a complement to obtain validity.

External validity concerns the extent to which the causal relationship is generalizable. The effects should exist even if the persons, settings or treatments were changed. (Cook et al. 2002, p. 83) Typical threats to external validity are interaction of treatment and selection, setting, and history (Cook et al. 2002, p. 87). An example of a thread for external validity in my study is that the tasks may not be realistic for customer service chat. Another is that the participants may be either too experienced or too inexperienced for the tasks.

4.2 Independent variables

When choosing independent variables, one must ensure that they match the key ideas of the research problem (Hornbæk et al. 2013). According to Gergle & Tan (2014), it is important to have a well-controlled variation, a clear operational definition,

and a meaningful range, in independent variables and levels. The conditions should be comparable, with a similar setting and equivalent functionality (Hornbæk et al. 2013).

There can be one or more independent variables in an experiment. Selecting more than one independent variable complicates the experiment structure and analysis of the results. However, having more independent variables may show interesting interactions between the conditions. In addition, an independent variable can have two or more levels that are compared. (Hornbæk et al. 2013)

Two independent variables were selected to study my research problem: layout of the chat agent user interface and the number of simultaneous chats. Two different values were chosen for both of the layouts. Thus, there were four different conditions in the experiment.

In the section 2.1, two most significant chat agent UI layouts were selected to be studied in this thesis. These two layouts were implemented for the experiment. In the first layout, the windowed layout, all of the chat windows were visible at the same time (Figure 5). In the second, the tabbed layout, only one chat window is visible at a time, and the other conversations are listed on the left side of the browser window (Figure 6).

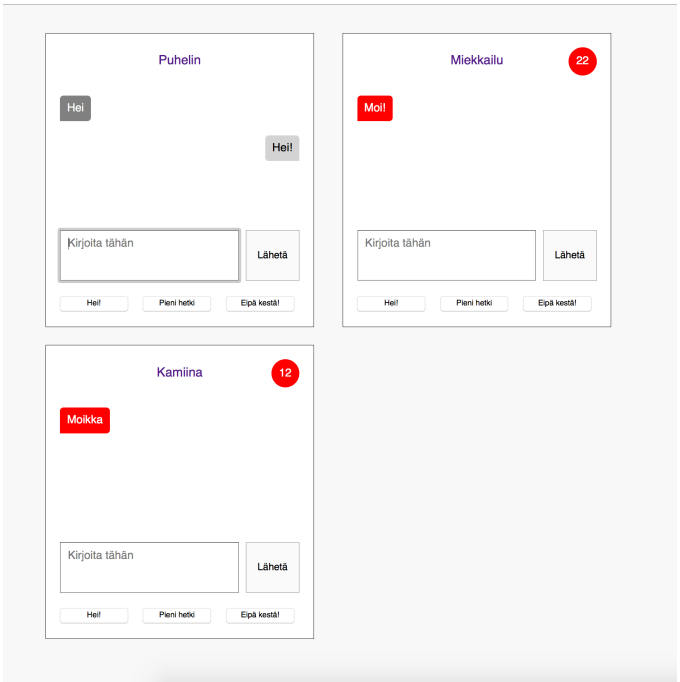


Figure 5: the windowed layout, where all of the chat windows are visible simultaneously.

The values for another independent variable - the amount of simultaneous chats - were decided based on previous studies and data from chat software. As introduced in section 2.1, the selected values were three and four.

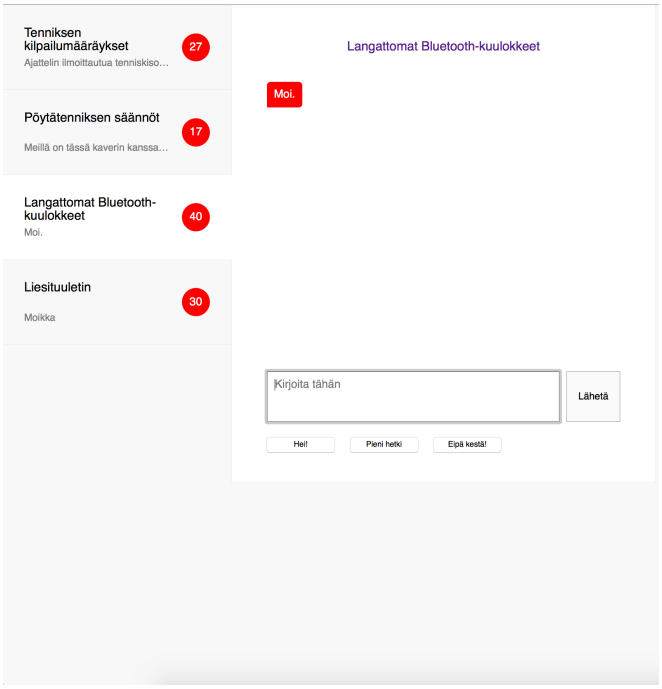


Figure 6: the tabbed layout, where all only one chat window is visible at a time.

4.3 Dependent variables

Dependent variables are chosen based on the formulated hypotheses, application domain, context, and previous work (Hornbæk et al. 2013). Dependent variables should be selected so that they can actually capture the effects that are being studied in the research problem (Gergle & Tan 2014). Collecting data for dependent variables can be quantitative, such as measuring task completion time or errors, or qualitative, such as interviewing user satisfaction. Typical dependent variables are, for example, task completion time, mental effort, learning, and quality of outcome (Hornbæk 2006).

What is difficult when selecting dependent variables is *operationalization* (Hornbæk et al. 2013). In this thesis, I measure usability. As stated in section 2.2, usability can be operationalized and measured in many different ways. I selected to measure it as efficiency, effectiveness, and satisfaction. Those were further operationalized into the following dependent variables: first response time, question response time, chat duration, accuracy, perception of efficiency, stress, control, frustration, and retention, and preference between layouts.

First response time and question response time were straightforward to measure. First response time was the time from receiving the first message in a chat to the answer of the participant. Question response time was measured as the time the participant used to give answer for the actual questions about the topic. Chat duration was measured as the time from the first message from the system in the chat to the last message from participant. Accuracy of the responses was measured as the number of errors in the questions.

Standard questionnaires, such as QUIS (Chin et al. 1988) or SUS (Brooke et al. 1996) would have been too long measure satisfaction in my study. Therefore, I used my own Likert-type rating scale questionnaires to measure the perception of efficiency, stress, control, frustration, and retention. The questionnaires contained five questions, rated in a scale from 1 to 5 (1="strongly disagree", 5="strongly agree"), and they were filled after each block in the experiment. Furthermore, participant's preference was measured with a final questionnaire after the whole experience. Both questionnaires are shown in Appendix A.

As a summary, all dependent variables and how they were measured, are shown in Table 1.

Category	Variable	How
Efficiency	First response time	Time from customer's first message to participant's first message in a chat
	Question response time	Time from customer's question about the topic of a chat to participant's answer
	Chat duration	Time from the first message of a chat to the last message of the chat
Effectiveness	Accuracy	Number of errors, i.e., wrong answers to the topic questions
Satisfaction	Perception of efficiency Perception of stress Perception of control Perception of frustration Perception of retention	Likert-type rating scale questionnaire after each block, containing five statements and a scale from 1 (strongly disagree) to 5 (strongly agree)
	Preference	Questionnaire after all blocks, asking which layout the participant preferred

Table 1: Dependent variables and how they are measured.

4.4 Experiment

4.4.1 Description

To study multitasking effects on customer service chat, I implemented a web-based chat prototype with ReactJS and Django. The basic idea of the experiment was to perform tasks in multiple simultaneous chats. To study response times and accuracy, the tasks had to contain questions to which the participants should answer. To avoid threats to external validity, experiments should be designed in a way that the tasks would be as close to real world tasks as possible (Hornbæk et al. 2013). Thus, I wanted to use conversational questions. Dresner & Barak (2006, 2009) used simulated conversations in their studies on conversational multitasking. The participants followed the conversations, and answered multiple-choice questions after

them. This approach would have been conversational, but it would have not measured the response time in a meaningful way.

Instead, I chose to ask something that the participants would have to answer synchronously, like in a real conversation. To test accuracy, the questions had to be designed in a way that there is a correct answer for each of them. To avoid threats to internal validity, the nature of the questions had to be something that the participants would not be likely to know in advance. Otherwise, the results would have depended on the general knowledge of the participants.

As a pre-study for this thesis I visited three giosg customer companies. During these visits, I observed a total of six customer service agents while they were working. In my observations, I was interested in how did they use the chat software and what kinds of tasks they performed with it. The general nature of the companies was quite different from each other, as well as their purpose for the chat. However, one thing that was common in all of them, was that the agents were more or less searching information for the customers they were serving.

From that observation, I came up with the idea of the task in my experiment. The task was to search information for "customers" in chats. Each chat had its own topic, for example a household appliance or a sports hobby. Questions about the topics were asked from the participants during chats. The questions about the household appliances were mostly something like "How wide is the washing machine?" or "Which program should I use to wash very dirty dishes?". Sports questions were something like "How long is the side of the football field?" or "I am at the level 14 in tennis, which level class should I participate in a competition?". All topics and questions used in the experiment are listed in Appendix B. In the experiment, answers for the questions were found from separate PDF (Portable Document Format) files. The files were user manuals for household appliances and rules for sports and sports competitions. Link to the PDF files is also in Appendix B.

The participants were guided not to give polite or full sentence answers for the questions. Instead, only one or a few words were sufficient. Politeness would have increased the customer satisfaction, but it would have been out of the scope of this study. Moreover, the participants were not customer service professionals. Politeness could also have led to significant differences in the other measures, as formulating answers could have been more natural for some participants than others. Thus, leaving politeness out of scope reduced external validity as the reliability of the study, but at the same time, it increased the internal validity by ensuring more similar answers and results.

Besides the topic questions, greetings and thanks were included in the conversation, in order to increase the feeling of reality. In addition, they were used to see if the user would react quickly to first messages while searching answers for other messages. As mentioned earlier, first response time is important factor when it comes to customer service chat efficiency. It gives the customers a feeling that they are being served and thus increases the customer satisfaction.

Velaro (2012) stated that the participants are usually impatient and do not want to wait for answers over one minute in web chats. Luo & Zhang (2013) argued that a slow response from an agent may make a customer abandon the chat. Moreover,

Lockwood (2017) found that preventing 'dead-air' time (i.e., responding too slowly) is important in CSC. Therefore, the participants were asked to answer the messages as fast as possible. If the system had been waiting the answer for too long, it could ask "Oletko vielä siellä?" ("Are you still there?"). This was an attempt to increase reality and interaction within conversations, like the greeting and thanks messages.

4.4.2 Design

As stated earlier, one or more independent variables can be manipulated in a study. If more than one independent variable is studied, a *factorial design* is used. In a factorial design, the effects of all the independent variables are examined at the same time. Moreover, their interaction with each other can be studied. (Cook et al. 2002, p. 263-264) This thesis investigates two independent variables, both with two possible values. Thus, my study uses a 2x2 factorial design.

Independent variables can be studied either with *between-subjects* design or *within-subjects* design. Between-subjects design means that the participants are divided into groups, and one value of each independent variable is assigned for each group. In a within-subjects design, each participant conducts the experiment with all of the values of the independent variable. In addition, mixed designs where some variables are studied between-subjects and other variables are studied within-subjects, are possible. In a within-subjects design it is easier to get statistically significant results with less participants. (Hornbæk et al. 2013) I used within-subjects design in this study.

In the experiment, each participant performed the tasks in four 8-minute task blocks. Every block had one of the four different conditions: the windowed layout with three simultaneous chats, the windowed layout with four simultaneous chats, the tabbed layout with three simultaneous chats, or the tabbed layout with four simultaneous chats. An example of the experiment blocks is shown in Figure 7.

According to Gergle & Tan (2014), one should be careful when deciding the order of the conditions for participants, because there is a possibility of learning during the experiment. Effects of learning on the results can be minimized by *counterbalancing* the conditions (Gergle & Tan 2014). This means that the order of the conditions is varied between the participants. I used Latin Square method (Bradley 1958) for counterbalancing. There were four conditions, that can be ordered in 24 different ways. Thus, I needed at least 24 or 48 participants to run the experiment with each different order of the conditions. The participants did not know the order in advance and could not predict it during the experiment.

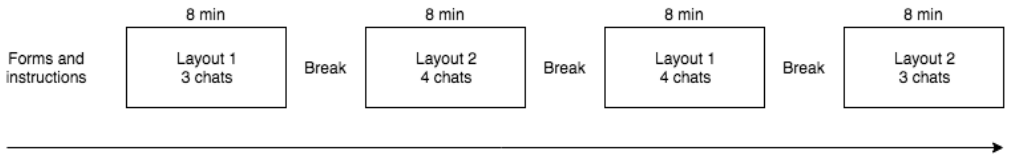


Figure 7: Example experiment blocks.

4.4.3 Participants

Other things to be decided when constructing the experiment are, who should participate in the experiment, and how many participants are needed. If the research focuses on a certain domain, it is reasonable to find only participants whose characteristics are sufficient to investigate the research problem (Purchase 2012, p. 76). These characteristics might include gender, age, or IT-skills, for example. Otherwise, the participants should be selected in a way that the results could be generalized to other people as well (Purchase 2012, p. 76). The number of participants depends on the design of the study. As mentioned earlier, within-subjects design needs fewer participants than between-subjects design, to be statistically significant. According to Hornbæk et al. (2013) a typical number of participants in HCI studies is 20. Having too few participants is not powerful enough to detect most effects.

24 participants (9 females, 15 males) were recruited to my experiment through email, Facebook posts and flyers. The age range of the participants was from 22 to 31 (M=25.83, SD=2.35). Because the whole experiment was in Finnish, all of the participants read and wrote fluent Finnish. In addition, all of them reported that they are using computer daily. Participants were compensated with a movie ticket for their participation.

Even though the focus of this study was in a customer service chat, I did not want the participants to be customer service specialists. One reason for this was that a customer service specialist might have been used to either one of the layouts, which could have been a threat to external validity. In addition, the experiment tested multitasking behavior instead of customer service skills.

4.4.4 Setting

The experiment was conducted in Aalto University's Department of Computer Science, in Computer Science building. A laboratory room was used, and there was a screen, a keyboard and a mouse connected to a laptop on a table (Figure 8). The web browser used in this study was Mozilla Firefox.

4.4.5 Procedure

In the beginning of each trial, the participant read the Information sheet for participant (Appendix C), and filled the Consent form (Appendix D) and the Basic information questionnaire (Appendix E). Then the participant read instructions for the actual task (Appendix F) from PDF slides on the computer. The instructions described the general nature of the task. It contained images of both of the layouts, as well as instructions on how to use them. In addition, the usage of the PDF manual was explained. All of the message types were also introduced, as well as how they should be answered.

As soon as the participant had read the instructions and had no questions about them, they started performing the actual task. Depending on the first condition, three or four chat windows were opened in the browser window. First conversation

Figure 8: The experiment setting.

started immediately and every 10 seconds one more conversation started, until all the chat windows had a conversation in progress.

There were four types of messages in a conversation: greeting at the beginning, wondering if the participant is still there, thanks in the end of the conversation and questions about the topic. For the first three types, there were answer buttons that the participant was informed to use to answer (Figure 9). For the actual questions about the topic, the participant was asked to use a free text area to write the message and a button to send the message (Figure 9).

All of the "customer" messages in the chats were automated. A random sample of chats were taken from giosg chat database, in order to create realistic wait times between messages. Because the blocks were only 8 minutes long, some of the wait times had to be decreased, so that there would be enough interaction during each block.

As mentioned in earlier, answers for the questions about the topic were found from the PDF manual. The PDF file was opened beside the chat windows by clicking the chat title (Figure 10). The file opened from the contents page. The questions were formed in a way that it was relatively easy to figure out from the contents page, from which part of the PDF the answers could be found. The contents page items were clickable, which made it easy to navigate to the right page.

Each topic contained from three to four questions about the topic, in addition to greeting and thanks. Once the thank was answered by the participant, the conversation ended and disappeared, and a conversation about a new topic opened into the same window.

After every 8-minute block, the participant filled a questionnaire about the block.

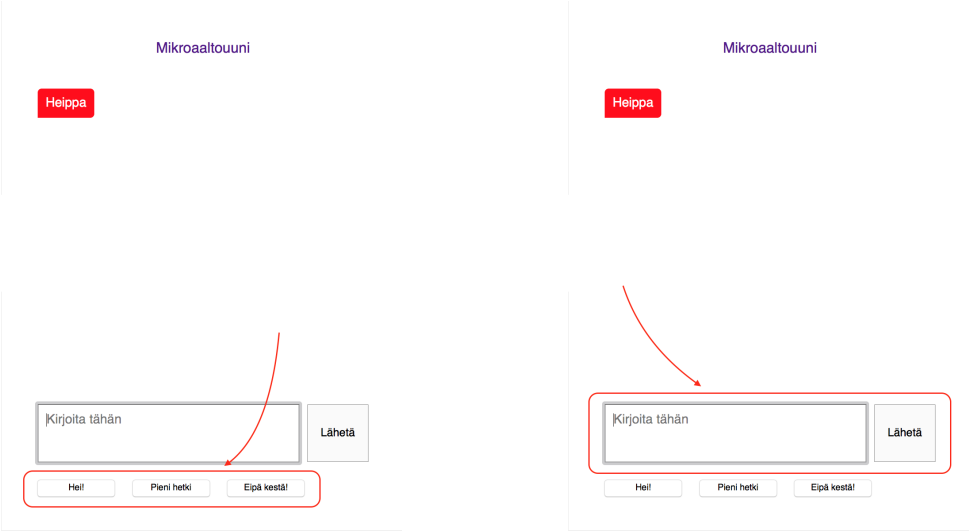


Figure 9: Buttons that are used to answer to other three message types than topic questions (left). Text area and send button are used to answer to topic questions (right).

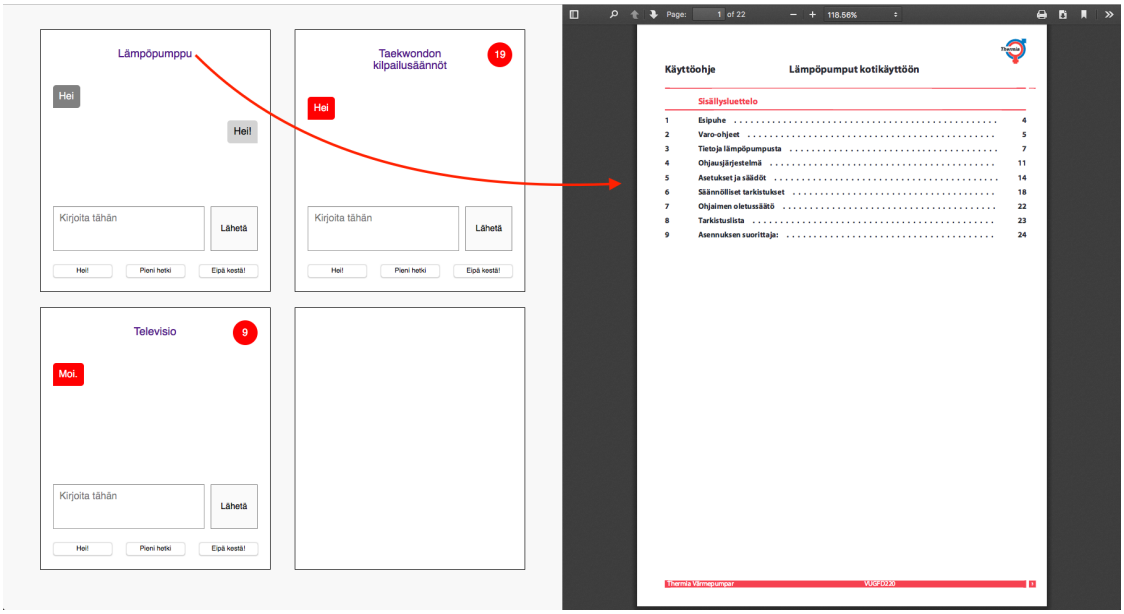


Figure 10: PDF is opened beside the chat windows by clicking a chat window title.

The questionnaire contained five questions about participant's feelings and how well they thought they were doing in the block.

After filling the questionnaire, the participant was informed to have a short break and to continue to the next block when they felt ready. When all of the four blocks were conducted, the participant was asked to fill in a final questionnaire. The final questionnaire had three questions: which layout the participant felt more pleasant to use, which layout the participant felt more efficient to use and whether they felt

difference between three and four simultaneous conversations.

4.5 Statistical analysis

Once the data from experiment is collected, *statistical analysis* aims to discover the possible effects of independent variables on dependent variables (Purchase 2012, p. 117). Different statistics (e.g., mean, median, standard deviation) are collected from large data sets, and the conditions are compared with each other with these data (Purchase 2012, p. 118).

However, the possible differences between conditions cannot be concluded to be the absolute truth. Because the experiment was conducted only with a sample of the population, the differences describe only a probability of the effect. Thus, *statistical significance* of the results has to be examined. (Purchase 2012, p. 125)

Statistical analysis methods are used to test a *null hypothesis*, that there is no difference between the conditions. The methods output a *p-value*, which tells the probability that the estimated values are similar (i.e., that the null hypothesis holds true). A typical threshold for p-value is 0.05, which I am using in this study as well. If p-value is below this threshold, it means that the null hypothesis is unreliable and it can be rejected. (Hornbæk et al. 2013)

There are two types of statistical methods, *parametric statistical methods* and *non-parametric statistical methods*. If the data is normally distributed, parametric methods are used, otherwise non-parametric methods are needed (Purchase 2012, p. 125). Statistical analyses for within-subjects designs are called *repeated measures* and analyses for between-subjects designs are called *independent measures* (Purchase 2012, p. 128). Different statistical methods are used depending on whether repeated measures are used, as well as whether there are two or more conditions in the experiment. Because my experiment used a within-subjects design, I introduce repeated measures statistical methods.

In case of two conditions, for normally distributed data, *t-test* can be used. *One-tailed* t-test can be used to test whether either of the conditions has a greater effect. *Two-tailed* t-test is used to test if there is any difference between the conditions or not. (Purchase 2012, p. 129) A non-parametric method for two conditions is *Wilcoxon test* (Purchase 2012, p. 142).

If there are more than two conditions for normally distributed data, *analysis of variance* (ANOVA) or *linear mixed models*, for example, can be used as statistical analysis methods. They determine if any condition has effect on dependent variables. An example of a non-parametric method for more than two conditions is *Friedman test* (Purchase 2012, p. 146). Another option is generalized linear mixed model, which is used like linear mixed model, but it assumes that the data is not normally distributed. Instead, a distribution describing the data has to be provided for the model.

There were four conditions in my experiment. The data for some of the dependent variables were normally distributed and for some were not (see Appendix G). One advantage in mixed models over ANOVA are that they handle random effects, i.e., variation that is caused by something else than the fixed conditions in the experiment.

They also handle missing values better than ANOVA. In my experiment, I assumed that there might be variation between participants. Therefore, I chose to use linear mixed model and generalized linear mixed model as statistical analysis methods for efficiency and effectiveness data in my study. Rating scale questionnaire data, I used Wilcoxon ranked sum test to study the effects of layout and amount of chats for each question independently. Wilcoxon ranked sum test assumes that the data is scaled and not paired.

5 Results

A total of 556 chats were initiated for the 24 participants during the experiment trials. For 537 conversations the participants sent at least one message, and 229 of the conversations were finished (answer for the last message was sent by the participant). Data for one condition (the tabbed layout with four chats) was missing from one participant.

The goal was to study whether there were any differences in the dependent variables between the windowed layout and the tabbed layout, and between three and four simultaneous chats. Moreover, interaction between layout and chat amount was investigated.

First response time and question response time were approximately gamma distributed (Figure G2, Figure G4). Thus, first response time and question response time were investigated with generalized linear mixed models with gamma distribution. Chat duration was approximately normally distributed (Figure G8). Therefore, it was investigated with a linear mixed model. Accuracy, which was calculated by number of errors, was approximately Poisson distributed (Figure G6), so I used a generalized linear mixed model with Poisson distribution, to analyze it.

I used layout and amount of chats as fixed effects for the models, and participant as a random effect. The following expression was used for all mixed models:

$$variable \sim as.factor(layout) * as.factor(chats) + (1|participant) \quad (1)$$

Satisfaction measures, that were rating scale data, were analyzed with Wilcoxon ranked sum tests. For each question, the layouts were compared with three and four chats independently, and then the amounts of chats were compared with the windowed layout and the tabbed layout. Threshold for significance was lowered to α=0.01, because the tests were consecutive.

This section contains subsections for the results of each variable. Finally, all hypotheses and their results are summarized.

5.1 First response time

Table 2 shows mean, median, standard deviation, minimum and maximum values for first response times in each condition. The results for generalized linear mixed model test for first response time are shown in Appendix H1. The results indicate that the first response time was faster in the windowed layout, and the difference was significant (t=2.731, p=0.00631). In addition, the first response time was significantly faster with three chats than with four chats (t=2.760, p=0.00578). Therefore, hypothesis H1 that first response time is faster for the windowed layout is supported, and hypothesis H5 that there is no significant difference in first response time between the chat amounts is not supported. The estimates for first response time in each condition are shown in Table 3. There was no interaction between layout and chat amount, which means that the results for layouts did not depend on chat amount, and vice versa. Box plot for first response time is shown in Figure 11. Intraclass correlation coefficient (ICC) (McGraw & Wong 1996) was also calculated from the results. For

first response time it was ICC=0.1812, which means that there were no individual differences among participants.

Condition		Mean	Median	Std	Min	Max
windowed layout	3 chats	7.21	2.37	13.2	1.24	98.9
	4 chats	9.64	2.98	14.8	1.28	97.3
tabbed layout	3 chats	9.86	3.10	13.9	1.54	75.8
	4 chats	11.7	3.77	20.9	1.61	199.0

Table 2: First response time mean, median, standard deviation, minimum and maximum values (in seconds).

	3 chats	4 chats
windowed layout	6.10	8.61
tabbed layout	8.66	12.22

Table 3: Generalized linear mixed model estimates for first response time (in seconds) in each condition.

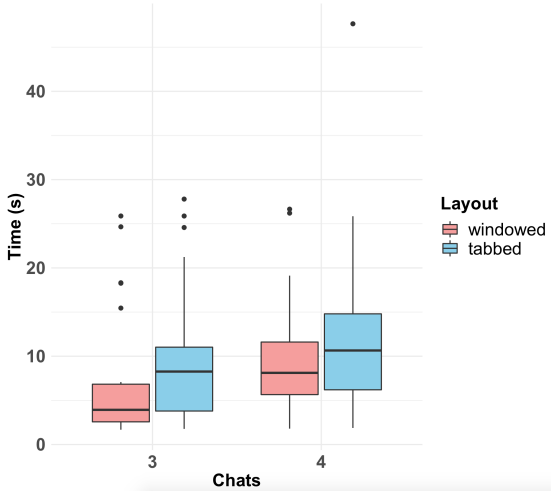


Figure 11: First response time box plot.

5.2 Question response time

Table 4 shows mean, median, standard deviation, minimum and maximum values for question response times in each condition. The results for generalized linear mixed model test with gamma distribution for question response time are shown in Appendix H2. The results show no difference between layouts. However, there was a significant difference between three and four simultaneous chats (t=8.237, p=<2e-16). Thus, hypotheses H2 that there is no difference in question response time between layouts, and H6 that the question response time is slower in four chats, are supported. The estimates for question response time in each condition are shown in Table 5. Again, no interaction was found between layout and chat amount. Box plot for question response time is shown in Figure 12. Intraclass correlation for question response time was even smaller than for first response time (ICC=0.0846).

Condition		Mean	Median	Std	Min	Max
windowed layout	3 chats	54.4	44.6	37.1	7.11	214.0
	4 chats	78.8	68.2	51.6	6.98	436.0
tabbed layout	3 chats	51.5	43	33.6	7.48	212.0
	4 chats	81.4	70.3	57.3	11.9	382.0

Table 4: Question response time mean, standard deviation, minimum and maximum values (in seconds) for each condition.

	3 chats	4 chats
windowed layout	54.88	80.23
tabbed layout	51.96	75.96

Table 5: Generalized linear mixed model estimates for question response time (in seconds) in each condition.

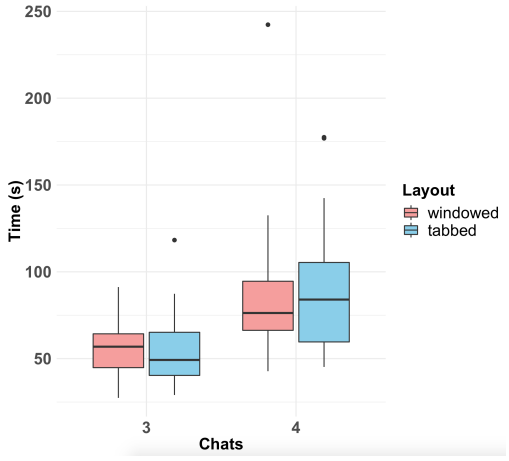


Figure 12: Question response time box plot.

5.3 Accuracy

Accuracy was measured as the number of errors. Mean, median, standard deviation, and minimum and maximum values for number of errors in each condition are shown in Table 6. Generalized linear mixed model results for number of errors are shown in Appendix H3. The results show no significant difference either between layouts or between number of simultaneous chats. Therefore, hypotheses H3 that accuracy is higher in the tabbed layout and H7 that the accuracy is higher in three simultaneous chats are not supported. In addition, there was no interaction between layout and chat amount. Box plot for number of errors is shown in Figure 13. Intraclass correlation for accuracy was ICC=0.1862.

Condition		Mean	Median	Std	Min	Max
windowed layout	3 chats	0.667	0.5	0.816	0	3
	4 chats	0.75	1	0.847	0	3
tabbed layout	3 chats	0.583	0	0.717	0	2
	4 chats	0.652	0	0.775	0	2

Table 6: Mean, median, standard deviation, minimum and maximum values for number of errors.

	3 chats	4 chats
windowed layout	0.60	0.67
tabbed layout	0.52	0.59

Table 7: Generalized linear mixed model estimates for number of errors in each condition.

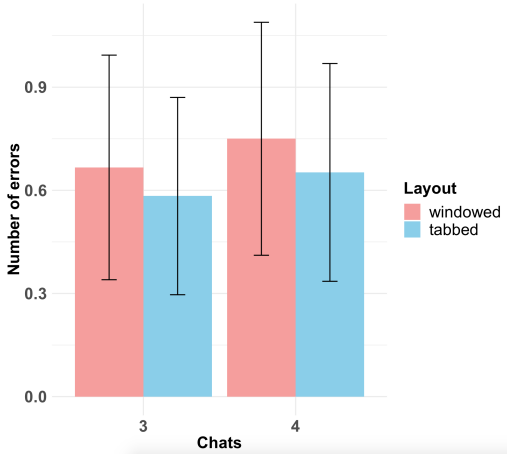


Figure 13: Number of errors box plot.

5.4 Chat duration

Table 8 shows mean, median, standard deviation, minimum and maximum values for chat durations in each condition. The results for linear mixed model test for chat duration are shown in Appendix H4. The linear mixed model results show that there was no significant difference in chat duration between the windowed layout and the tabbed layout. The results show that the chat duration with four simultaneous conversations was higher than with three conversations, and the difference was significant (t=3.074, p=0.00239). Thus, hypotheses H4 that chat duration does not differ between the windowed layout and 2, and H7 that chat duration is higher with four chats, are supported by the results. The estimates for chat duration in each condition are shown in Table 9. There was no interaction between layout and number of chats. Box plot for chat duration is shown in Figure 14. Intraclass correlation was small also for chat duration (ICC=0.1594).

Condition		Mean	Median	Std	Min	Max
windowed layout	3 chats	323.0	321.0	79.0	155.0	463.0
	4 chats	360.0	370.0	70.5	170.0	473.0
tabbed layout	3 chats	323.0	319.0	68.2	221.0	472.0
	4 chats	368.0	391.0	67.9	245.0	464.0

Table 8: Chat duration mean, median, standard deviation, minimum and maximum values (in seconds) for each condition.

	3 chats	4 chats
windowed layout	325.38	363.22
tabbed layout	325.25	363.09

Table 9: Linear mixed model estimates for chat duration (in seconds) in each condition.

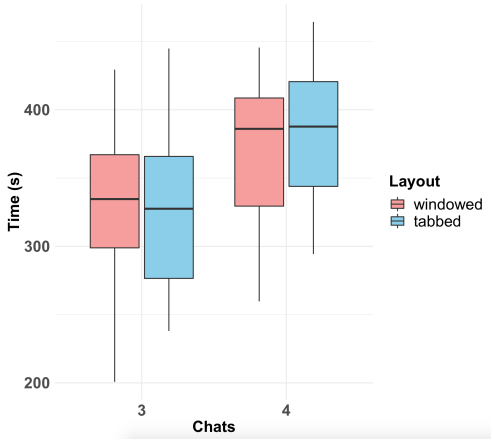


Figure 14: Chat duration box plot.

5.5 Satisfaction

Satisfaction was measured with Likert-type rating scale questionnaires after each task block. This subsection shows the results of rating frequencies, as well as Wilcoxon test results, for each satisfaction variable in each condition. As a summary, the means for each satisfaction variable are shown in Figure 15.

5.5.1 Efficiency

Table 10 shows ratings for item "Sain tehtävät tehokkaasti suoritettua tässä osiossa" ("I got the tasks done efficiently in this block") (1="strongly disagree", 5="strongly agree"). Wilcoxon test results do not show any significant differences between the layouts or the chat amounts.

Condition		1	2	3	4	5
windowed layout	3 chats	0	1	5	13	5
	4 chats	0	3	8	11	2
tabbed layout	3 chats	0	2	5	9	8
	4 chats	0	5	4	10	4

Table 10: Frequencies of efficiency ratings.

5.5.2 Stress

Table 11 shows ratings for item "Tämä osio oli stressaava" ("This block was stressful"). Wilcoxon test results indicated that there was no significant difference between the layouts with either chat amount. However, for the windowed layout, there was a significant difference between three and four chats (W=137.5, p=0.001209). For the tabbed layout, there was a cue for difference between chat amounts, but with α=0.01, it was not significant (W=185, p=0.04207). The results indicate that when chat amount is constant, the stress level is similar with both layouts. However, if the chat amount is increased, it might cause more stress with the windowed layout. Therefore, hypothesis H9 that the tabbed layout causes less stress is supported.

Condition		1	2	3	4	5
windowed layout	3 chats	6	14	3	1	0
	4 chats	3	4	11	6	0
tabbed layout	3 chats	5	10	8	1	0
	4 chats	1	9	7	5	1

Table 11: Frequencies of stress ratings.

5.5.3 Control

Table 12 shows ratings for item "Tunsin hallitsevani tilanteen hyvin tässä osiossa" ("I felt I had the situation well under control in this block"). Wilcoxon test results do

not show any difference between the layouts with either amount of chats. Moreover, the difference between chat amounts with either of the layouts was not significant.

Condition		1	2	3	4	5
windowed layout	3 chats	0	2	5	13	4
	4 chats	0	5	8	9	2
tabbed layout	3 chats	0	2	3	15	4
	4 chats	2	3	7	8	3

Table 12: Frequencies of control ratings.

5.5.4 Frustration

Table 13 shows ratings for item "Tunsin itseni turhautuneeksi tässä osiossa" ("I felt frustrated in this block"). The Wilcoxon test results indicate that there was no difference between layouts with either amount of chats. With the windowed layout, there was a cue for difference between three and four chats, but with α=0.01 it was not significant (W=192, p=0.03612). With the tabbed layout, there was no difference between the chat amounts.

Condition		1	2	3	4	5
windowed layout	3 chats	13	8	3	0	0
	4 chats	8	7	2	6	1
tabbed layout	3 chats	9	11	4	0	0
	4 chats	7	9	3	4	0

Table 13: Frequencies of frustration ratings.

5.5.5 Retention

Table 14 shows ratings for item "Tässä osiossa oli helppo muistaa keskusteluiden aiheet/sisällöt" ("In this section, it was easy to remember the topics/contents of the conversations"). The Wilcoxon test results do not show any differences between the layouts or the chat amounts.

Condition		1	2	3	4	5
windowed layout	3 chats	0	2	7	11	4
	4 chats	1	7	4	8	4
tabbed layout	3 chats	0	6	5	5	8
	4 chats	1	4	9	5	4

Table 14: Frequencies of retention ratings.

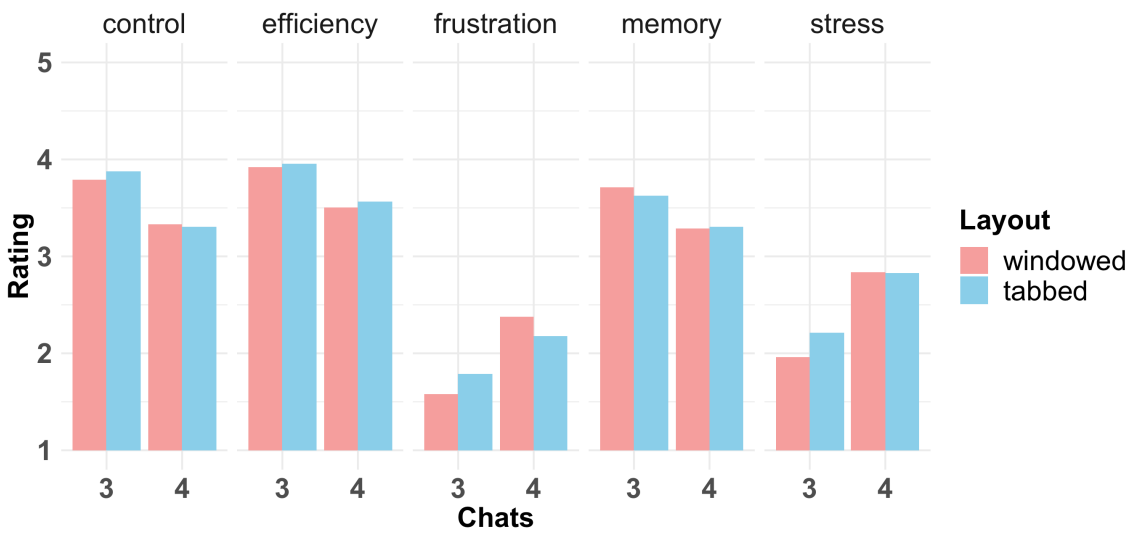


Figure 15: The means for each satisfaction variable in each condition.

5.5.6 Preference

As the result for final questionnaire after the experiment, exactly half of the participants (12) preferred the the windowed layout, whereas another half preferred the the tabbed layout. Closer investigation of the results reveals a significant correlation between the layout in the first block of the experiment and the preferred layout. Pearson correlation test showed a negative correlation between preferred layout and the layout in the first block of the experiment (r=-0.5, p=0.01285, t=-2.708, df=22). This means that there is a probability that the first layout was not the preferred one after the experiment.

5.6 Summary

As a summary, all hypotheses and their results are shown in Table 15.

Hypothesis		Result
H1	First response time is faster in the windowed layout.	Supported
H2	There is no significant difference in question response time between the windowed layout and the tabbed layout.	Supported
H3	Accuracy is higher in the tabbed layout.	Not supported
H4	There is no significant difference in chat duration between the windowed layout and the tabbed layout.	Supported
H5	There is no significant difference in first response time between three and four simultaneous chats.	Not supported
H6	Question response time is slower with four simultaneous chats.	Supported
H7	Accuracy is higher with three simultaneous chats.	Not supported
H8	Chat duration is higher with four simultaneous chats.	Supported
H9	the tabbed layout causes less stress than the windowed layout.	Supported

Table 15: Summary of hypotheses and their results.

6 Discussion

The goal of this thesis was to study the effects of chat layout on usability in customer service chat multitasking. To achieve the goal, usability was studied in an experiment with two different chat agent user interface layouts, and with two different amounts of simultaneous chats.

This section first discusses the findings and implications of the study. Second, validity and limitations of the study are discussed. Based on them, ideas for future studies are proposed.

6.1 Findings and implications

First response time, question response time, chat duration, accuracy, and satisfaction were measured in the experiment. I predicted the the windowed layout to be more efficient in terms of first response time, but the the tabbed layout to be more effective with higher accuracy. Moreover, I assumed that efficiency and effectiveness would be higher with three simultaneous conversations than with four simultaneous conversations.

According to the results, first response time was significantly faster in the windowed layout. Switching to other chats does not require extra clicks in the windowed layout. In addition, switching chat in the tabbed layout hides the interrupted chat and makes resuming for it more difficult. Therefore, switching between chats is an easier task in the windowed layout, and the participants may have a lower threshold to do it. Moreover, as Catanzaro et al. (2006) found in their study comparing tiled and tabbed chat layouts, new messages were easier to detect when all the chat windows are shown simultaneously. This is potentially part of the reason for faster first response time in the windowed layout in my study.

In my experiment, first response was an easy task, because it was about answering to a greeting with an answer button. Another easy task in the experiment was the last response. However, last response time did not differ between layouts. I assume that the participants did not feel as much pressure for answering the last message than other messages. Questions about the topics were not easy to answer, because the participants had to find the answers from the manuals. However, if there would have been some other easy questions during the chats, I assume that their responses would have been faster in the windowed layout also.

The results show that even if the stress level did not differ significantly between the layouts with three chats, with four chats it increased significantly in the windowed layout but not in the tabbed layout. As mentioned in section 3, Jeuris & Bardram (2016) found that if other tasks are visible, it may increase perceived time pressure and cause stress. In addition, Mark et al. (2008) found that interruptions speed up working, but on the other hand, they cause more stress. My results for higher stress level and faster first response time in the windowed layout are in line with these findings. In the windowed layout, new messages are more salient than in the tabbed layout, which causes the bottom-up visual attention to be driven to them more easily. Therefore, a possible explanation for higher stress level in the windowed

layout is that it causes more interruptions than the tabbed layout.

The rest of the usability measures did not differ significantly between the layouts. Question response time, chat duration, and accuracy were similar in the layouts. In addition, other satisfaction measures than perception of stress did not differ significantly between the layouts either.

There are many potential factors affecting the question response time in both layouts: Switching time is slower in the tabbed layout. Moreover, the interrupted chat is visible during interruption in the windowed layout, which may decrease the resumption time. On the other hand, in the windowed layout the other chats are more salient in the windowed layout, which potentially drives visual attention away from the current chat, slowing down the response time in the current chat.

Even though the first response time was faster in the windowed layout, the chat duration did not differ between the layouts. Faster first response time in the windowed layout indicates more switching between chats. Therefore, the time saved in faster first responses is probably lost in interruptions caused by the switching, because resuming after interruption takes time. However, to confirm the actual reasons for similar question response times and chat durations, further studies should be done. Those are discussed more in the next subsection.

I predicted accuracy to be higher in the tabbed layout, but no difference was found between the layouts. I assumed that more frequent switching in the windowed layout would result in more interruptions. As stated earlier, errors are more likely after interruptions. However, no difference in accuracy was found between the layouts. One potential reason for this is that, as Salvucci et al. (2009) suggested, visibility of an interrupted task rehearses it in memory, and therefore it may not take as much time to resume a chat in the windowed layout. Thus, the activation of the interrupted chat does not decay as much as in the tabbed layout, and it is more probably retrieved from memory as the most active goal. Other possible reason is that the tasks were not sufficient to measure accuracy.

The results show significant differences in some aspects of usability between three and four simultaneous chats. Question response time was faster with three chats, and chat duration was longer with four chats. With four chats, there were more interrupting notifications and more tasks to interleave and interact with, which probably was the reason for longer response times and duration for individual chats.

I assumed first response time to be similar with both chat amounts, but it was slower with four simultaneous chats. The reason is probably the same as for question response time and chat duration, that there were more chats to interact with.

Accuracy was predicted to be higher in three simultaneous chats, but the results show no difference between the amounts. With four chats, there should be more interruptions than with three chats, so the amount of interruptions seemed not to affect accuracy. In addition, even though the probability to retrieve a wrong goal from memory after interruption was higher with four chats, it did not cause more errors. Thus, it seems that either there is no difference in accuracy between the studied layouts or amounts of simultaneous chats, or as reckoned above, the tasks may not have been sufficient to measure it in a chat environment.

Other satisfaction measures than stress did not differ significantly between the

layouts nor between the chat amounts. This indicates that the participants did not feel subjective differences between the layouts, which is in line with the question response time, chat duration, and accuracy results. However, even though the efficiency was decreased when the amount of chats was increased, the participants did not perceive it. My assumption is that they did not feel that they were responding slower, because they were responding more chats at a time.

The results of this study indicate that there are no significant differences on usability between the layouts. This is also supported by the result that half of the participants preferred each layout. Therefore, either of the layouts can be suggested to be used in customer service chat. For giosg, the results indicate that using the the windowed layout is reasonable, because it may lower the first response time. However, the number of visible chat windows should be considered carefully, to avoid critical increase in stress level. I suggest that the number of visible windows could be limited, showing only the most active chats. Nevertheless, more research is needed, to be able to conclude the amount of visible chat windows. The next subsection discusses the limitations of this study and the ideas for future research.

6.2 Limitations and future ideas

In the experiment, there was given a possibility to reply "Pieni hetki" ("Just a moment") with an answer button. The participants used it quite differently in the tests. Some of them used it only to respond to the "Oletko vielä siellä?" ("Are you still there?") messages, but some of them used it to react to almost every message, before they had time to actually answer them. This may have decreased the internal validity of the study. In fact, I measured how many times "Pieni hetki" was used as a first reaction to messages in each condition. There were no significant differences between conditions. However, the intraclass correlation was much higher than in other measures (ICC=0.58). It means that there were differences between participants. Therefore, in future studies it should be either excluded or the instructions on how to use it should be clearer.

As I found in the literature review, customers are willing to wait for high quality answers. Accuracy, which was used to measure effectiveness in the tasks in this study, did not give a sufficient understanding on that aspect. Therefore, also end-user satisfaction could be included to the measures in the future, with more realistic chat tasks.

More realistic tasks would also give a better understanding on the interaction behavior in chat multitasking. Part of the messages should be something that the participants would be able to answer straight away, without searching for an answer. It is possible that in my study, there was no realistic interleaving between chats, because the participants knew that there would not be any easier questions waiting in the other conversations. As stated above, I assume that faster first responses in the windowed layout may indicate faster responses for easy questions in general, in the windowed layout. More realistic conversations would test this assumption.

In addition, interaction behavior should be further studied with eye-tracking and mouse-tracking. Those would give an insight into whether interaction behavior in

the different layouts would actually differ from each other. Those studies could also further justify and explain the results of this study.

As discovered in the literature review, memory plays a significant role in multitasking. In my study, memory was involved in a way that in interruption situations, the participants had to remember what information they were looking for before the interruption. However, they did not actually have to remember the earlier messages in the conversations (except in a couple questions which were related to the previous question). In real conversations, it is more important to remember the contexts and contents of the conversations, to be able to continue after interruptions. Thus, in the future, more realistic conversations would show the effects of memory in multitasking situation more clearly. This might affect question response time and chat duration.

Finally, according to the results, exactly half of the participants preferred the windowed layout and the other half preferred the tabbed layout. A negative correlation was found between the layout in the first block of the experiment and the preferred layout. This means that the participants may have been confused in the beginning, before learning how the experiment is actually conducted. In the future, there could be a short training for both layouts before the actual experiment, in addition to the instructions.

7 Conclusions

In this thesis, I studied how different chat user interface layouts affect usability in customer service chat multitasking. The effects were studied with an experiment, where two different layouts and two amounts of simultaneous chats were tested by measuring usability as first response time, question response time, chat duration, accuracy, and user satisfaction. In the first, windowed layout, all simultaneous chats were visible for the user at the same time in separate chat windows. In the second, tabbed layout, only one chat was shown for the user at a time, and it was changed from the left side of the user interface. The amounts of simultaneous chats studied in the experiment were three and four.

The windowed layout showed clear advantages on faster first response time. As answering to the first message was an easy task in the experiment, it may be that answering to any easy question would be faster in that layout. However, this layout was found to cause more stress for the users, when chat amount was increased from three to four, whereas the tabbed layout did not show a significant effect on that. Response time and chat duration were considerably higher with four simultaneous chats in both layouts. These results give validity for the experiment measures. No effects on accuracy between the two layouts or the chat amounts were found in the study.

As an answer to the main research problem, this study shows that chat layout may improve efficiency by speeding up response times for easy questions, such as greetings, but with a cost of increased stress level. However, fast responses to those messages indicate more interruptions during chats, and therefore, the time saved is lost. If first reactions are wanted to be provided fast, layout should show all the simultaneous chats in separate windows. On the other hand, if the amount of simultaneous chats is particularly high, showing only one chat at a time for the user could prevent a higher increase in the stress level. Increasing the amount of simultaneous chats affects usability by decreasing efficiency.

In the future, more realistic conversation tasks should be used to study the effects on retention with the layouts. Because of the constantly increasing demand for customer service chat, further studies should be done with even higher amounts of simultaneous chats. The accuracy measured in this study potentially gave no sufficient understanding on effectiveness, and thus, end-user satisfaction should be included to the measures in the future research. Moreover, this study could be complemented with eye-tracking and mouse-tracking, to give better understanding on how users are actually interacting with the interfaces, and on what the users spend the time. With those studies, the results of this study could be justified and explained further, and further decisions could be made for choosing and designing the layout.

Moreover, at giosg, if we decide to keep the windowed layout, we should do further user research with the new user interface designs. Especially, we should measure the stress levels and other satisfaction measures, in addition to response times and chat duration.

This thesis provides a valuable groundwork for further studies on chat multitasking

research. There was no previous research actually studying customer service chat usability from the user interface point-of-view. This study also provides data for future modeling of conversational multitasking. The findings of this thesis are in line with the existing multitasking research. First, the findings support the argument that interruptions and switching between tasks may increase efficiency to a certain point. Second, faster task completion times due to interruptions come with a cost of increased stress levels. Last, the findings suggest that resuming after an interruption takes time.

References

Adamczyk, P. D. & Bailey, B. P. (2004), If Not Now, When?: The Effects of Interruption at Different Moments Within Task Execution, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, pp. 271–278.

Adler, R. F. & Benbunan-Fich, R. (2012), 'Juggling on a high wire: Multitasking effects on performance', *International Journal of Human-Computer Studies* **70**(2), 156–168.

Altmann, E. M. & Trafton, J. G. (2002), 'Memory for goals: an activation-based model', *Cognitive science* **26**(1), 39–83.

Anderson, J. R. (1993), *Rules of the Mind*, Psychology Press.

Aral, S., Brynjolfsson, E. & Van Alstyne, M. (2006), Information, Technology and Information Worker Productivity: Task Level Evidence, *in* 'Proceedings of the International Conference on Information Systems', AIS Electronic Library, pp. 285–306.

Bailey, B. P. & Iqbal, S. T. (2008), 'Understanding Changes in Mental Workload during Execution of Goal-Directed Tasks and Its Application for Interruption Management', *ACM Transactions on Computer-Human Interaction (TOCHI)* **14**(4), 21.

Bannister, F. & Remenyi, D. (2009), 'Multitasking: the uncertain impact of technology on knowledge workers and managers', *The Electronic Journal Information Systems Evaluation* **12**(1), 1–12.

Bardhi, F., Rohm, A. J. & Sultan, F. (2010), 'Tuning in and tuning out: media multitasking among young consumers', *Journal of Consumer Behaviour* **9**(4), 316–332.

Bevan, N. (1995), 'Measuring usability as quality of use', *Software Quality Journal* **4**(2), 115–130.

BoldChat (2015), 'Live Chat Performance Benchmarks 2015', https://www.bold360.com/~/media/457e059b899648348c92e4a64ac41505.pdf. Accessed: 2018-04-16.

Borst, J. P., Taatgen, N. A. & van Rijn, H. (2015), What Makes Interruptions Disruptive? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption, *in* 'Proceedings of the 33rd annual ACM conference on human factors in computing systems', ACM, pp. 2971–2980.

Bowman, L. L., Levine, L. E., Waite, B. M. & Gendron, M. (2010), 'Can students really multitask? An experimental study of instant messaging while reading', *Computers & Education* **54**(4), 927–931.

Bradley, J. V. (1958), 'Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design', *Journal of the American Statistical Association* **53**(282), 525–528.

Brooke, J. et al. (1996), 'SUS – A quick and dirty usability scale', *Usability evaluation in industry* **189**(194), 4–7.

Catanzaro, J. M., Risser, M. R., Gwynne, J. W. & Manes, D. I. (2006), Military Situation Awareness: Facilitating Critical Event Detection in Chat, *in* 'Proceedings of the Human Factors and Ergonomics Society Annual Meeting', Vol. 50, SAGE Publications Sage CA: Los Angeles, CA, pp. 560–564.

Chin, J. P., Diehl, V. A. & Norman, K. L. (1988), Development of an instrument measuring user satisfaction of the human-computer interface, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, pp. 213–218.

Clarkson, D. (2010), 'Making Proactive Chat Work', http://www.roosit.nl/files/en/documents/Forrester_making_proactive_chat_work.pdf. Accessed: 2018-04-16.

Comm100 (2018), 'Live Chat Benchmark Report 2018', http://www.ec3.co.za/uploads/2/6/3/7/26378480/comm100-live-chat-benchmark-report-2018.pdf. Accessed: 2018-04-11.

Comm100 Network Corporation (2018), 'Live chat', https://www.comm100.com/livechat/. Accessed: 2018-06-15.

Cook, T. D., Campbell, D. T. & Shadish, W. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Boston, Boston, MA.

Cutrell, E. B., Czerwinski, M. & Horvitz, E. (2000), Effects of Instant Messaging Interruptions on Computing Tasks, *in* 'CHI'00 extended abstracts on Human factors in computing systems', ACM, pp. 99–100.

Czerwinski, M., Cutrell, E. & Horvitz, E. (2000), Instant Messaging: Effects of Relevance and Timing, *in* 'People and computers XIV: Proceedings of HCI', Vol. 2, pp. 71–76.

David, P., Xu, L., Srivastava, J. & Kim, J.-H. (2013), 'Media multitasking between two conversational tasks', *Computers in Human Behavior* **29**(4), 1657–1663.

Dindar, M. & Akbulut, Y. (2016), 'Effects of multitasking on retention and topic interest', *Learning and Instruction* **41**, 94–105.

Dresner, E. & Barak, S. (2006), 'Conversational Multitasking in Interactive Written Discourse as a Communication Competence', *Communication Reports* **19**(1), 70–78.

Dresner, E. & Barak, S. (2009), 'Effects of visual spatial structure on textual conversational multitasking', *Communication Quarterly* **57**(1), 104–115.

Duggan, G. B., Johnson, H. & Sørli, P. (2013), 'Interleaving Tasks to Improve Performance: Users Maximise the Marginal Rate of Return', *International Journal of Human-Computer Studies* **71**(5), 533–550.

eDigitalResearch (2014), 'Customer Service Benchmark', http://www.edigitalresearch.cowww.edigitalresearch.com/pdf/sample-benchmarks/Customer%20Service%20Benchmark%20March%202014.pdf. Accessed: 2018-04-11.

Ellis, Y., Daniels, B. & Jauregui, A. (2010), 'The effect of multitasking on the grade performance of business students', *Research in Higher Education Journal* **8**(1), 1–10.

Elmorshidy, A. (2013), 'Applying The Technology Acceptance And Service Quality Models To Live Customer Support Chat For E-Commerce Websites', *Journal of Applied Business Research* **29**(2), 589–595.

Gergle, D. & Tan, D. S. (2014), Experimental research in HCI, *in* 'Ways of Knowing in HCI', Springer, pp. 191–227.

Gillie, T. & Broadbent, D. (1989), 'What makes interruptions disruptive? A study of length, similarity, and complexity', *Psychological research* **50**(4), 243–250.

giosg.com Ltd (2018*a*), 'About giosg', https://www.giosg.com/about. Accessed: 2018-08-20.

giosg.com Ltd (2018*b*), 'Available Chat Features', https://www.giosg.com/download-the-full-features-list. Accessed: 2018-07-10.

giosg.com Ltd (2018*c*), 'giosg Live Chat', https://www.giosg.com/features/chat. Accessed: 2018-07-12.

González, V. M. & Mark, G. (2004), "Constant, Constant, Multi-tasking Craziness": Managing Multiple Working Spheres, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, pp. 113–120.

Hart, S. G. & Staveland, L. E. (1988), Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *in* 'Advances in psychology', Vol. 52, Elsevier, pp. 139–183.

Hembrooke, H. & Gay, G. (2003), 'The Laptop and the Lecture: The Effects of Multitasking in Learning Environments', *Journal of computing in higher education* **15**(1), 46–64.

Hornbæk, K. (2006), 'Current practice in measuring usability: Challenges to usability studies and research', *International journal of human-computer studies* **64**(2), 79–102.

Hornbæk, K. et al. (2013), 'Some Whys and Hows of Experiments in Human–Computer Interaction', *Foundations and Trends® in Human–Computer Interaction* **5**(4), 299–373.

Intercom (2018), 'Live chat', https://www.intercom.com/live-chat. Accessed: 2018-06-15.

ISO (2010), ISO 9241-210:2010 Ergonomics of human system interaction – Part 210: Human-centred design for interactive systems, Standard, International Standardization Organization (ISO), Switzerland.

Itti, L. & Koch, C. (2001), 'Computational modelling of visual attention', *Nature reviews neuroscience* **2**(3), 194–203.

Jeuris, S. & Bardram, J. (2016), 'Dedicated workspaces: Faster resumption times and reduced cognitive load in sequential multitasking', *Computers in Human Behavior* **62**, 404–414.

Kang, E. (2013), 'The Ideal Online Experience: What it Takes for Consumers to Click, Not Abandon', https://www.liveperson.com/connected-customer/posts/ideal-online-experience-what-it-takes-consumers-click-not-abandon. Accessed: 2018-08-02.

Kang, L., Wang, X., Tan, C.-H. & Zhao, J. L. (2015), 'Understanding the Antecedents and Consequences of Live Chat Use in Electronic Markets', *Journal of Organizational Computing and Electronic Commerce* **25**(2), 117–139.

Katidioti, I., Borst, J. P., van Vugt, M. K. & Taatgen, N. A. (2016), 'Interrupt me: External interruptions are less disruptive than self-interruptions', *Computers in Human Behavior* **63**, 906–915.

Kayako (2017), 'Live Chat Statistics', https://www.kayako.com/live-chat-software/statistics. Accessed: 2018-04-11.

Kazmi, S. H. A., Abid, M. M. et al. (2016), Online Purchase Intentions in E-Commerce, *in* 'Proceedings of 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)', Vol. 2, IEEE, pp. 570–573.

Kulbytė, T. (2018), '6 Customer Service Reports That Every Manager Needs'. Accessed: 2018-08-02.

Lee, J., Lin, L. & Robertson, T. (2012), 'The impact of media multitasking on learning', *Learning, Media and Technology* **37**(1), 94–104.

Li, S. Y., Blandford, A., Cairns, P. & Young, R. M. (2008), 'The Effect of Interruptions on Postcompletion and Other Procedural Errors: An Account Based on the Activation-Based Goal Memory Model', *Journal of Experimental Psychology: Applied* **14**(4), 314.

Likert, R. (1932), 'A technique for the measurement of attitudes', *Archives of psychology* **22**(140).

LiveChat Inc (2018), 'LiveChat', https://www.livechatinc.com/. Accessed: 2018-06-15.

LivePerson Inc (2018), 'LiveEngage', https://www.liveperson.com/liveengage/messaging. Accessed: 2018-06-15.

Lockwood, J. (2017), 'An analysis of web-chat in an outsourced customer service account in the Philippines', *English for Specific Purposes* **47**, 26–39.

Luo, J. & Zhang, J. (2013), 'Staffing and Control of Instant Messaging Contact Centers', *Operations Research* **61**(2), 328–343.

Mark, G., Gudith, D. & Klocke, U. (2008), The cost of interrupted work: more speed and stress, *in* 'Proceedings of the SIGCHI conference on Human Factors in Computing Systems', ACM, pp. 107–110.

McGraw, K. O. & Wong, S. P. (1996), 'Forming inferences about some intraclass correlation coefficients', *Psychological methods* **1**(1), 30–46.

McLean, G. & Osei-Frimpong, K. (2017), 'Examining satisfaction with the experience during a live chat service encounter-implications for website providers', *Computers in Human Behavior* **76**, 494–508.

Monk, C. A., Trafton, J. G. & Boehm-Davis, D. A. (2008), 'The Effect of Interruption Duration and Demand on Resuming Suspended Goals', *Journal of Experimental Psychology: Applied* **14**(4), 299.

Paul, C. L., Komlodi, A. & Lutters, W. (2015), 'Interruptive notifications in support of task management', *International Journal of Human-Computer Studies* **79**, 20–34. Integrating Knowledge of Multitasking and Interruptions Across Different Perspectives and Research Methods.

Purchase, H. C. (2012), *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*, 1st edn, Cambridge University Press, New York, NY, USA.

Rosenholtz, R., Li, Y. & Nakano, L. (2007), 'Measuring visual clutter', *Journal of vision* **7**(2), 1–22.

Salvucci, D. D. (2001), 'Predicting the effects of in-car interface use on driver performance: An integrated model approach', *International Journal of Human-Computer Studies* **55**(1), 85–107.

Salvucci, D. D. & Taatgen, N. A. (2008), 'Threaded Cognition: An Integrated Theory of Concurrent Multitasking', *Psychological review* **115**(1), 101.

Salvucci, D. D., Taatgen, N. A. & Borst, J. P. (2009), Toward a Unified Theory of the Multitasking Continuum: From Concurrent Performance to Task Switching, Interruption, and Resumption, *in* 'Proceedings of the SIGCHI conference on human factors in computing systems', ACM, pp. 1819–1828.

Sanbonmatsu, D. M., Strayer, D. L., Medeiros-Ward, N. & Watson, J. M. (2013), 'Who Multi-Tasks and Why? Multi-Tasking Ability, Perceived Multi-Tasking Ability, Impulsivity, and Sensation Seeking', *PloS one* **8**(1), e54402.

Shackel, B. (1991), 'Usability – Context, framework, definition, design and evaluation', *Human factors for informatics usability* pp. 21–37.

Shae, Z.-Y., Garg, D., Bhose, R., Mukherjee, R. & Guven, S. (2007), Efficient Internet Chat Services for Help Desk Agents, *in* 'Services Computing, 2007. SCC 2007. IEEE International Conference on', IEEE, pp. 589–596.

Steele, I. (2017), 'Do Your Live Chat Agents Measure Up? The 9 Best Key Performance Indicators and How To Use Them'. Accessed: 2018-04-17.

Strayer, D. L. & Johnston, W. A. (2001), 'Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone', *Psychological science* **12**(6), 462–466.

TELUS International (2015), 'Best Practices: Online Chat Sales', http://web2.telusinternational.com/hubfs/Best_Practices_Online_Chat_Sales_0715.pdf?t=1461244896505. Accessed: 2018-04-11.

Tezcan, T. & Zhang, J. (2014), 'Routing and Staffing in Customer Service Chat Systems with Impatient Customers', *Operations research* **62**(4), 943–956.

Velaro (2012), 'How Long Will You Wait On Hold for Customer Service?', http://www.prweb.com/releases/2012/10/prweb9964730.htm. Accessed: 2018-07-25.

Vergic (2018), 'Vergic Engage Platform', https://www.vergic.com/. Accessed: 2018-06-15.

Wang, Y., Echenique, A., Shelton, M. & Mark, G. (2013), A Comparative Evaluation of Multiple Chat Stream Interfaces for Information-intensive Environments, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, pp. 2717–2720.

Warr, A., Chi, E. H., Harris, H., Kuscher, A., Chen, J., Flack, R. & Jitkoff, N. (2016), Window Shopping: A Study of Desktop Window Switching, *in* 'Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems', ACM, pp. 3335–3338.

Wasserman, L. (2001), 'Live interaction: What's needed on the web', *Customer Interaction Solutions* **19**(7), 58–61.

Wickens, C. D. (1984), Processing resources in attention, *in* 'Varieties of Attention', New York: Academic Press, pp. 63–101.

Wickens, C. D. (2002), 'Multiple resources and performance prediction', *Theoretical issues in ergonomics science* **3**(2), 159–177.

Wickens, C. D., Gutzwiller, R. S. & Santamaria, A. (2015), 'Discrete task switching in overload: A meta-analyses and a model', *International Journal of Human-Computer Studies* **79**, 79–84.

Wolfe, J. M. & Horowitz, T. S. (2004), 'What attributes guide the deployment of visual attention and how do they do it?', *Nature reviews neuroscience* **5**(6), 495–501.

Yan, S., Tran, C. C., Chen, Y., Tan, K. & Habiyaremye, J. L. (2017), 'Effect of user interface layout on the operators' mental workload in emergency operating procedures in nuclear power plants', *Nuclear Engineering and Design* **322**, 266–276.

Zendesk (2015), 'Zendesk Benchmark: Live Chat Drives Highest Customer Satisfaction', https://www.zendesk.com/company/press/zendesk-benchmark-live-chat-drives-highest-customer-satisfaction/. Accessed: 2018-04-11.

Zendesk (2018), 'Zendesk chat', https://www.zendesk.com/chat/. Accessed: 2018-06-15.

A Questionnaires

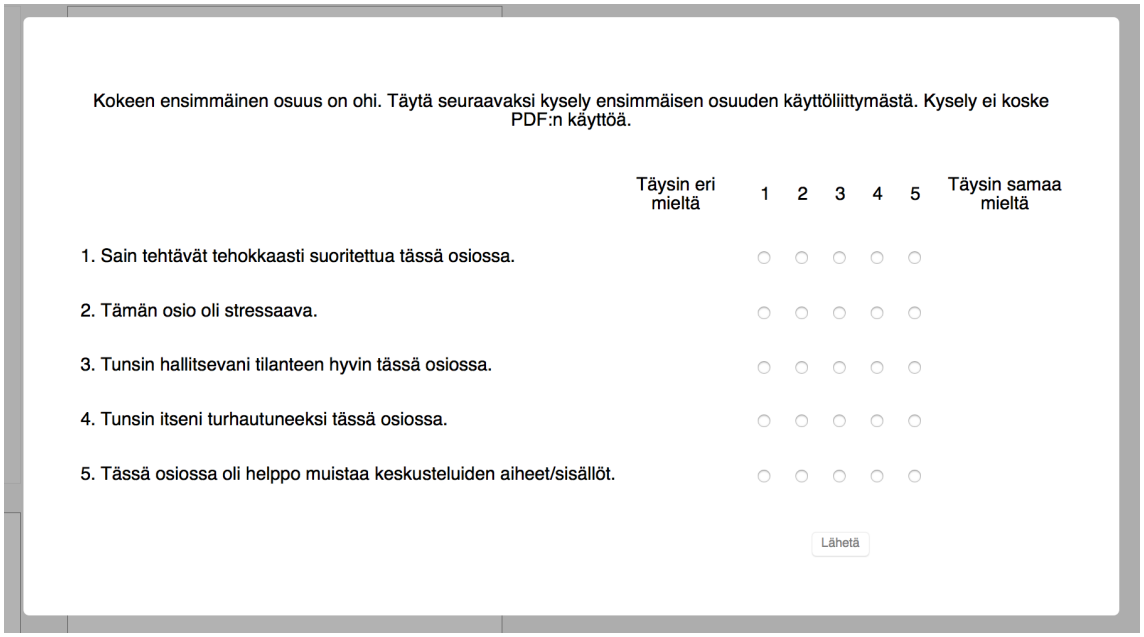


Figure A1: Questionnaire for rating perception of efficiency, stress, control, frustration and retention after each task block.

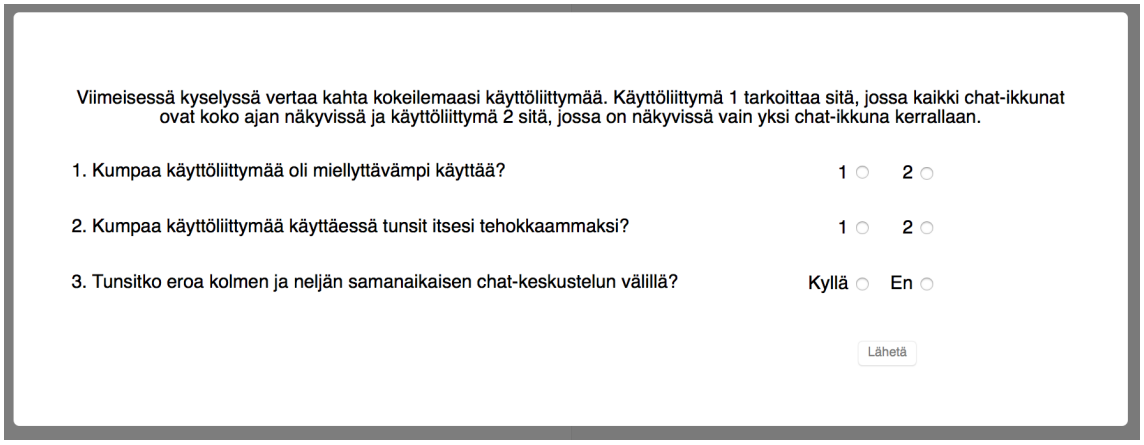


Figure A2: Final questionnaire for rating preference and feeling of efficiency between layouts, as well as feeling of difference between three and four simultaneous chats.

B Topics and questions

The PDF files for the manuals can be found from Google drive: https://drive.google.com/drive/folders/1Ht_xBN856J2ubmMt6dpZyeT3dRSQxue7?usp=sharing.

Topic	Question	Answer
Agility	Olen osallistumassa ensimmäistä kertaa agilitykilpailuun. Koirani säkäkorkeus on 40cm, mihin kilpailuluokkaan (koko) minun pitää osallistua?	Medi(M)
	Kuinka monta eri tasoluokkaa on olemassa?	3
	Onko agilitykilpailun tarkoituksena yhtenä osaalueena koiran ulkonäön arviointi?	Ei
Arkkupakastin	Tarvitsisin vastauksen muutamaan kysymykseen, jotta tiedän ostanko arkkupakastimen. Kuinka leveä pakastin on?	595 mm
	Sitten kysymys asennuksesta. Voinko asentaa pakastimen varastoon, jossa on talvella +5 astetta lämmintä?	Et
	Entä mikä on alhaisin lämpötila, jolle pakastimen voi säätää käyttöpaneelista?	-24°C
Astianpesukone	Ostin astianpesukoneen, ja minulla olisi muutama kysymys siihen liittyen. Mikä ohjelma on tarkoitettu erittäin likaisten astioiden pesuun (ohjelman numero)?	3
	Selvä, mitä lämpötilaa tämä ohjelma käyttää?	70°C
	Sitten vielä yksi kysymys koneen asetuksista. Asuinpaikassani veden kovuus on 32°dH, mihin asentoon vedenpehmennin tulee siis asettaa?	7
Cheerleadingin kilpailusäännöt	Joukkueemme on osallistumassa cheerleadingin SM-kisoihin ja haluaisin selvittää muutaman asian kilpailusäännöistä. Kuinka suuri kilpailualue on SM-kisoissa?	14x16 m
	Entä kuinka monta ulkoista spotteria joukkueella saa olla kilpailussa enintään?	6
	Saako ulkoinen spotteri tukea huojuvaa pyramidia?	Ei saa
Generaattori	Missä tätä generaattoria on tarkoitus käyttää?	Asuntovaunuissa, matkailuautoissa, kaupallisessa käytössä olevissa ajoneuvoissa
	Kuinka paljon se painaa?	96,5kg

Generaattori	Mitä lisävarusteita generaattoriin on saatavana?	Ulkoinen tiivistesarja
	Minkälainen huolto pitäisi tehdä generaattorin käytön ensimmäisenä kuukautena?	Vaihdattaa öljy
Grilli	Mietin grillin ostoa ja olisi pari kysymystä siitä. Mitkä ovat grillauspinta-alan pituus ja leveys?	Pituus 49.6 cm, levetys 90.0 cm
	Kuinka pitkä takuu grillillä on?	2 vuotta
	Entä mitä lisävarusteita grilliin on saatavana?	Suojahuppu ja elektroninen grillivarras
Jalkapallo-säännöt	Ajattelin pitää tietovisan eri urheilulajeista ja mietin tässä jalkapallokysymyksiä, tarvitsisin niihin vastaukset. Ensimmäiseksi, mikä on pelikentän sivurajan minimipituus?	90 m
	Mikä on pallon ympärysmitta?	68-70cm
	Kuinka monta pelaajaa joukkueessa saa olla enintään?	11
	Sitten vielä yksi. Kuinka kaukana vastapuolen pelaajien pitää olla kulmakaaresta kulmapotkutilanteessa?	9.15 m
Jääkaappi	Tarvitsisin apua jääkaapin asennuksessa. Kuinka kauas sähköliedestä jääkaappi tulee vähintään sijoittaa?	3 cm päähän
	Olen kytkemässä jääkaappia toimintaan, laitoin juuri pistokkeen seinään ja nyt kuuluu joku hälytysääni. Mitä teen?	Paina lämpötilavalitsinta 1
	Jääkaapista kuuluu välillä naksahduksia. Kuuluvatko ne normaaleihin käyntiääniin?	Kyllä
	Voisin vielä varmuuden vuoksi ottaa talteen huoltopalvelun puhelinnumeron, mikä se on?	0207510700
Kamiina	Voinko käyttää ostamani kamiinan polttoaineena sytytysnestettä?	Et
	Entä mitä materiaalia keittolevy on?	Valurautaa
	Okei, sitten vielä yksi. Kuinka pitkä takuu kamiinalla on?	12 kk
Kiuas	Minulla olisi pari kysymystä koskien kiuastanne. Millä etäisyydellä kiukaan pitää olla sen takana olevasta seinästä?	100 mm
	Mikä on pisin aika, jonka päähän kiukaan voi ajastaa lämpenemään?	8 tuntia
	Saako löylyvetenä käyttää merivettä?	Ei
Koripallon pelisäännöt	Olen perustamassa koripallojoukkuetta ja olisi muutama kysymys. Kuinka monta pelaajaa joukkueessa voi enintään olla?	12
	Mitkä ovat pelikentän pituus ja leveys?	Pituus 28 m, leveys 15 m
	Kuinka monta erää ottelussa pelataan?	4
	Entä kuinka monta minuuttia yksi erä kestää?	10 min

Kuivauskaappi	Kuinka korkea kuivauskaappi on?	1900 mm
	Kuinka kauan lingotun pyykin kuivaus pikakuivauksella kestää?	noin 120 min
	Mitä pesuainetta kuivauskaapin puhdistukseen voi käyttää?	Mietoa saippualiuosta
Kuivausrumpu	Haluaisin kysyä pari kysymystä kuivausrummusta ennen ostopäätöstä. Kuinka leveä kuivausrumpu on?	600 mm (950mm luukun ollessa auki)
	Mikä on täysin kuivaksi kuivaavan ohjelman maksimitäyttömäärä?	8 kg
	Kuuluuko ajastettu kuivaus laitteen lisätoimintoihin?	Kyllä
	Entä onko laitteessa ohjelmaa villavaatteille?	Kyllä
Langattomat Bluetooth-kuulokkeet	Ostin juuri uudet Bluetooth-kuulokkeet ja tarvitsisin vastauksen muutamaan kysymykseen. Mitä kaikkea pakkauksen pitäisi sisältää?	Kuulokkeet, usb-latauskaapeli ja pikaopas
	OK, minulta näyttäisi puuttuvan ohjeet ja siksi tässä kyselen. Yritin muodostaa pariliitosta, mutta se kysyy jotain salasanaa. Mikä se mahtaa olla?	0000
	Kuinka kauan kuulokkeiden lataaminen pitäisi yleensä kestää?	Noin 2 tuntia
	Kuinka suuri kuulokkeiden toimintasäde on?	10 m
Liesituuletin	Olen käyttänyt liesituuletinta jonkin aikaa ja mieleeni heräsi muutamia kysymyksiä siitä. Miten saan intensiivitehon päälle? Laitteessani on ohjauspaneeli 1.	Paina +, kun tuuletin on teholla 3
	Kuinka usein rasvasuodatin pitäisi puhdistaa?	2 kk välein
	Entä mitä puhdistusainetta voi käyttää laitteen alumiini- ja muoviosien puhdistukseen?	Lasinpesuainetta
Lämpöpumppu	Lämpöpumppuni ohjausjärjestelmän vihreä merkkivalo alkoi vilkkumaan eilen, mitä se tarkoittaa?	Jokin hälytys on aktiivisena
	Mikä asetus pitää olla päällä, jotta ohjausjärjestelmä sallii vain sähkövastuksen toiminnan?	Lisälämpö
	Entä mikä on oletussäätö lämpötilalle, jos otan lämmityksen pois?	17°C
Mankeli	Olen katsellut mankelia mutta tarvitsisin vähän lisätietoja, jotta tiedän ostanko sen. Mikä on mankelin syvyys?	314 mm
	Entä mankeliliinan leveys?	55 cm
	Mikä on lakanoiden mankeiloimisaika?	2-4 min
Miekkailu	Olen miettinyt miekkailun aloittamista, mutta en tiedä mikä kolmesta miekkailulajista pitäisi valita. Mikä kolmesta miekkatyypistä (kalpa, floretti, säilä) saa olla painavin?	Kalpa
	Kuinka paljon se saa enintään painaa?	770 g
	Sitten vielä kiinnostaisi osuma-alueet. Mikä on säilämiekkailun osuma-alue?	Koko vartalo vyötäröstä ylöspäin
	Entä kalpamiekkailun?	Koko vartalo

Mikroaaltouuni	Minulla olisi pari kysymystä viime viikolla ostamastani mikrosta. Luulen että siinä on jokin häiriö, koska näytössä palaa kolme nollaa. Mikä voisi olla syy tähän?	Sähkökatko
	Selvä. Entä mikä on suurin teho jonka voin valita?	800 W
	Mitä tällä teholla tulisi kuumentaa?	Nesteitä
	Olen kuullut, että metalliastiaa ei saisi laittaa mikroon, mistä se johtuu?	Mikroaallot eivät läpäise metallia
Nelikopteri	Kiinnostaisi tietää muutama juttu tästä nelikopterista. Kuinka painava kopteri on?	1380 g
	Entä mikä on sen maksiminopeus?	20 m/s
	Kuuluko kotiinpaluutoimintoon kotiin paluu, jos akku on vähissä?	Kyllä
	Kuinka pitkä takuu kopterilla on?	1 vuosi
Ompelukone	Mikä on tässä ompelukoneessa ompelen enimmäisleveys?	5 mm
	Entä mille välille ompelen pituuden voi asettaa siksak-ompelussa?	0.5 - 4
	Kuinka painava kone on?	5 kg
Painepesuri	Mikä on mallin P 130.2 nimellispaine?	12 MPa/120 bar
	Entä nimellisvirrankulutus?	2.300 kW
	Kuinka pitkä takuu laitteessa on?	2 vuotta
Pyykinpesukone	Olen miettinyt pesukoneen ostoa, ja haluasin kysyä pari juttua ennen päätöstä. Mikä on koneen leveys?	600 mm
	Entä maksimitäyttömäärä?	7 kg
	Voiko koneen lisätoimintoihin kuuluvalla ajastuksella ajastaa koneen käynnistymään vuorokauden päästä?	Ei voi
Päältäajettava ruohonleikkuri	Haluaisin vertailla kahta ruohonleikkurimallia, LM2148 M ja LM2153 MD. Kummassa näistä on suurempi teho?	LM2153 MD
	Entä onko niiden polttoainesäiliöiden tilavuuksissa eroa?	Ei
	Mitkä niiden leikkuuleveydet ovat?	48 mm ja 53 mm
Pöytätenniksen säännöt	Meillä on tässä kaverin kanssa erimielisyyksiä pöytätenniksen säännöistä. Mikä on pöydän pituus?	274 cm
	Kuinka korkealle pallon kuuluu nousta pelaajan kämmenestä syötössä?	16 cm
	Entä kuinka kuinka monta pistettä pelaajan pitää vähintään saada, jotta voi voittaa erän?	11
Taekwondon kilpailusäännöt	Lapseni on osallistumassa ensimmäiseen taekwondokilpailuun ja muutama asia on vielä epäselvä. Mihin kilpailuluokkaan hänen pitää osallistua? Hän on 9-vuotias.	Värivyöt, C-juniorit
	Käydäänkö tässä luokassa pyykkipoikaottelua?	Kyllä
	Mitä punnituksessa pitää olla päällä?	T-paita ja dobokin housut

Televisio	Yritän muodostaa internet-yhteyttä ostamaani televisioon, mutta se ei tunnu onnistuvan. Mikä on minimi verkkoyhteyden nopeus, jotta internet-yhteys voidaan muodostaa?	10 Mbps
	Kytkin USB-näppäimistön televisioon, millä painikkeella pääsen palaamaan valikossa edelliseen näyttöön?	ESC-näppäimellä
	Entä mitä Windows-näppäin tekee?	Näyttää television asetukset
Tenniksen kilpailumääräykset	Ajattelin ilmoittautua tenniskisoihin, mutta edellisistä on niin kauan aikaa, että haluaisin tarkistaa pari juttua. Olen tasolla 14, mihin tasoluokkaan minun kuuluu osallistua?	C
	Olen vähän huolissani omasta jaksamisestani, kuinka pitkä tauko kahden normaalipituisen saman luokan ottelussa on vähintään?	60 min
	Kuinka montaa palloa tennisotteluissa voi korkeintaan käyttää?	6
Uuni	Harkitsen tässä uunin ostoa, mutta haluaisin ensin tietää laitteen mitoista vähän. Mikä on leveimmän keittolevyn halkaisija?	180 mm
	Entä kuinka monta kannatintasoa uunissa on?	4
	Sitten vielä pari kysymystä uunin asennuksesta. Voiko uunin asentaa 200 mm päähän seinästä?	Kyllä
	Hyvä, entä kuinka paljon tyhjää tilaa lieden päällä pitää vähintään olla?	650 mm

Table B1: Topics and questions used in experiment tasks.

C Information sheet

Tietotekniikan laitos, Perustieteiden korkeakoulu, Aalto-yliopisto

TIEDOKSI TUTKIMUKSEEN OSALLISTUVALLE

“Diplomityön käyttöliittymätutkimus”

Tutkimuksen nimi ja aihe: Käyttöliittymän ja samanaikaisten chat-keskustelujen määrän vaikutus monisuoritukseen: millainen käyttöliittymä tukee tehokasta interaktiota usean samanaikaisen chat-keskustelun välillä.

Tutkimusmenetelmän kuvaus: Tämä on kontrolloitu koe. Kokeessa keskitytään suorituskykyyn ja kokemukseen koehenkilön suorittaessa tehtäviä eri käyttöliittymillä. Koe koostuu neljästä tehtäväosiosta, kunkin osion jälkeen täytetään kysely kokemuksesta ja pidetään tauko. Tehtävien prosessit ja tulokset mitataan.

Tutkimuksen tarkoitus: Tutkimuksen tarkoitus on selvittää, kuinka erilaiset käyttöliittymät ja eri määrät samanaikaisia chat-keskusteluja vaikuttavat monisuoritukseen selainpohjaisessa asiakaspalveluchatissa. Tutkimus on osa diplomityötä. Henkilötietojesi käsittely on tarpeen yleisen edun vuoksi tieteellistä tutkimusta ja akateemista ilmaisua varten.

Rahoitus ja vastaava tutkija: European Research Council. Vastaava tutkija on tutkijatohtori Jussi Jokinen (jussi.jokinen@aalto.fi, +358 45 1961429).

Aika: Koe kestää noin 60 minuuttia.

Soveltuvuus tutkimukseen: Seuraavat kriteerit vaaditaan: sujuva suomen kielen lukeminen ja kirjoitus, sujuva tietokoneen käyttö, täysi-ikäisyys, normaali tai silmälaseilla/piilolinsseillä normaaliksi korjattu näkö, ei tietokoneen käyttöön vaikuttavia kognitiivisia tai liikunnallisia häiriöitä.

Korvaus: Finnkinon elokuvalippu

Vapaaehtoinen osallistuminen: Tutkimukseen osallistuminen on vapaaehtoista. Sinulla on oikeus keskeyttää osallistumisesi missä tahansa vaiheessa syytä ilmoittamatta, ilman seurauksia.

Tutkimukseen osallistuvan oikeudet: Tutkimukseen osallistuvalla on seuraavat tietosuojalaissa määrätyt oikeudet: 1) oikeus päästä tarkastamaan omat henkilötiedot, 2) oikeus oikaista tiedot, 3) oikeus vastustaa henkilötietojen käsittelyä, 4) oikeus tulla unohdetuksi eli oikeus tietojen poistamiseen.

Edellä mainituista oikeuksista tullaan mahdollisesti poikkeamaan, jos tutkimuksella on yleisen edun mukaiset tarkoitukset ja tutkimukseen osallistuvan oikeudet todennäköisesti estävät tarkoitusten saavuttamisen tai vaikeuttavat sitä suuresti ja tällaiset poikkeukset ovat tarpeen näiden tarkoitusten täyttämiseksi. Oikeuksiesi laajuus on sidottu henkilötietojesi käsittelyperusteeseen ja voimassaolevaan lainsäädäntöön.

Mahdolliset riskit ja niiden ennaltaehkäisy: Väsymys ja stressi. Ehkäisemme näitä jakamalla kokeen neljään osioon, joiden väleissä on tauot. Vaikka koetilanne saattaa tuntua intensiiviseltä, siitä on turha ottaa paineita, sillä kokeessa ei testata koehenkilöä vaan käyttöliittymää.

Tutkimushenkilökunnan kanssa kommunikointi kokeen aikana: Pyri esittämään kokeeseen tai osallistumiseesi liittyvät kysymykset joko ennen koetta tai osioiden välisillä tauoilla.

Tutkimustilanteen kuvaus: Saat tutkimuksen alussa ohjeet kokeen suorittamiseen. Kokeen aikana mitataan suorituskykyä. Saadulla datalla ei arvioida sinua, vaan vertaillaan erilaisia käyttöliittymiä. Kunkin tehtäväosion jälkeen mielipiteitäsi ja kokemuksiasi kysytään kysymyslomakkeella.

Datan kerääminen: 1) Suorituskyvyn mittarit: nopeus, virheet, tehtävien suoritusaika; 2) Kysely-data: mielipiteet ja kokemukset; 3) Henkilökohtaiset tiedot: nimi, sähköpostiosoite, sukupuoli, ikä, kokemus tietokoneen käytöstä. Henkilökohtaiset tiedot kerätään ainoastaan koehenkilön kanssa kommunikointia varten sekä ikä- ja sukupuolijakauman raportointiin tutkimuksen tuloksissa.

Tiedonsiirto EU:n ulkopuolelle: Tietojasi ei siirretä EU:n ulkopuolelle.

Nimettömyys, turvallinen varastointi, luottamuksellisuus: Tutkimuksen dataa tullaan käyttämään ainoastaan tieteelliseen tarkoitukseen ja se pidetään salassa. Kaikki data anonymisoidaan. Mitään vihjeitä identiteetistäsi ei jää tietokantaan tallennettuun dataan. Kaikki data tallennetaan turvallisesti ja luottamuksellisesti ja on saatavilla vain tutkimuksen suorittajalle (Jenni Pajukoski). Kaikki henkilökohtainen informaatio hävitetään, kun sitä ei enää tarvita.

Vakuutuksen kattavuus: Aalto-yliopiston vakuutus korvaa kokeen aikana sattuvat vahingot ja onnettomuudet.

Yhteystiedot:

Aalto-yliopisto on rekisterinpitäjä tässä tutkimuksessa.

Tutkimukseen liittyvissä kysymyksissä voit olla yhteydessä:

1) Tutkimuksen suorittajaan: Jenni Pajukoski, jenni.pajukoski@aalto.fi, +358 41 5077170
2) Tutkimuksen vastaavaan tutkijaan: Jussi Jokinen, jussi.jokinen@aalto.fi, +358 45 1961429

Voit olla yhteydessä Aalto-yliopiston tietosuojavastaavaan, jos sinulla on kysyttävää henkilötietojen käsittelystä ja suojauksesta: Jari Söderström, tietosuojavastaava@aalto.fi, 09 47001.

Jos koet, että henkilötietojasi on käsitelty tietosuojalainsäädännön vastaisesti, sinulla on oikeus tehdä valitus valvontaviranomaiselle, tietosuojavaltuutetulle (lue lisää: http://www.tietosuoja.fi).

Jos suostut osallistumaan tähän tutkimukseen, allekirjoita suostumuslomake.

D Consent form

Diplomityön käyttöliittymätutkimus

SUOSTUMUSLOMAKE

Minä ……. suostun osallistumaan Jenni Pajukosken diplomityön käyttöliittymätutkimukseen.

Olen lukenut ja ymmärtänyt minulle annetun, tutkimusta koskevan informaatiolomakkeen.

Ymmärrän, että kaikki data kerätään ainoastaan tieteelliseen tarkoitukseen. Tutkimuksen tarkoitus ja luonne on selostettu minulle kirjallisesti. Minulla on riittävästi tietoa tutkimuksen prosessista.

Ymmärrän, että osallistuminen tutkimukseen on täysin vapaaehtoista ja että minulla on oikeus keskeyttää osallistumiseni missä tahansa vaiheessa ilman seurauksia.

Annan luvan tietojeni tallentamiseen selostetulla tavalla.

Minulle on kerrottu, että nimetty tutkija antaa pyynnöstä lisätietoja tutkimuksen yleisistä periaatteista ja sen edistymisestä, tai minua koskevista tuloksista.

Ymmärrän, että anonymiteetti varmistetaan salaamalla identiteettini. Minulle on kerrottu, keitä ovat tutkimuksessa mukana olevat eri osapuolet, joilla on pääsy tietoihini. Ymmärrän tietojen säilyttämistä, suojaamista ja käyttöä koskevat käytännöt.

Tiedän, että kerättyjä tietoja ei luovuteta kolmansille osapuolille ilman kirjallista suostumustani. Tulosten kaikenlainen kaupallinen hyödyntäminen on kielletty.

Ymmärrän, että täysin anonyymi osa datasta voidaan luovuttaa muille tutkimusryhmille, jos annan siihen luvan.

(Valitse yksi seuraavista:)

[ ] Hyväksyn anonyymien otteiden luovutuksen.

[ ] Hyväksyn anonyymien otteiden luovutuksen vain, jos minulle kerrotaan kyseessä olevat tutkimusryhmät. Minulle on kerrottu, mitkä nämä otteet ovat.

[ ] En hyväksy otteiden luovutusta.

Allekirjoituksellani vahvistan osallistumiseni tähän tutkimukseen ja suostun olemaan vapaaehtoinen koehenkilö.

Päiväys ……...

TUTKIMUKSEEN OSALLISTUVA	TUTKIJA
Allekirjoitus……	Allekirjoitus……

Paikka.……...

E Basic information questionnaire

Diplomityön käyttöliittymätutkimus

ESITIETOLOMAKE

ID: _____

Ikä: _____

Sukupuoli:

[ ] nainen
[ ] mies
[ ] muu/en halua ilmoittaa

Käytän tietokonetta:

[ ] päivittäin
[ ] useammin kuin kerran viikossa
[ ] kerran viikossa
[ ] 2-3 kertaa kuukaudessa
[ ] kerran kuukaudessa
[ ] harvemmin kuin kerran kuukaudessa

F Experiment instructions

Ohjeet kokeen suorittamiseen

Kokeen tarkoitus

- Tutkimuksen kohteena on prototyyppi asiakaspalvelu-chatista, jossa käyttäjä käy useita chat-keskusteluja samanaikaisesti
- Kokeen tarkoitus on tutkia, kuinka erilaiset käyttöliittymät vaikuttavat monisuoritukseen usean samanaikaisen chat-keskustelun aikana

Kokeen rakenne

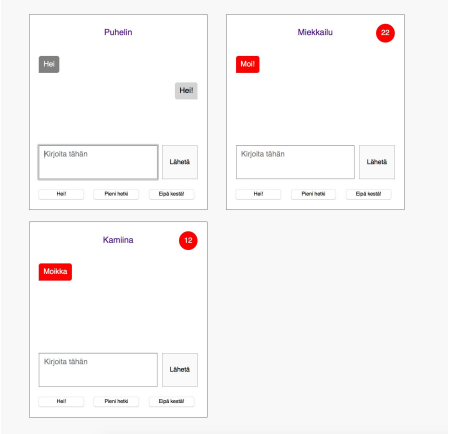
- Koe koostuu neljästä osiosta, joissa vaihtelevat erilaiset käyttöliittymät ja samanaikaisten chat-keskustelujen määrä
- Yksi osio kestää 8 minuuttia
- Kunkin osion jälkeen täytetään kysely osiosta ja pidetään pieni tauko

Tehtävän kuvaus

- Tehtävässä käydään useaa samanaikaista chat-keskustelua
- Kuhunkin keskusteluun liittyy jokin aihe, esimerkiksi kodinkone tai urheilulaji
- Keskustelussa esitetään aiheeseen liittyviä kysymyksiä, joihin löytyy vastaukset erillisestä ohjekirjasta (tästä lisää myöhemmin)
- Tavoitteena on vastata mahdollisimman moneen kysymykseen (oikein) annetussa ajassa

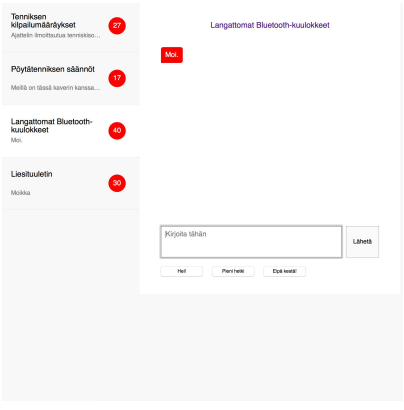
Käyttöliittymät

- Tässä käyttöliittymässä kaikki chat-ikkunat ovat näkyvissä koko ajan
- Ikkunan oikeassa ylänurkassa näkyy sekunteina aika, jonka verran asiakas on odottanut vastausta



Käyttöliittymät

- Tässä käyttöliittymässä näkyy kerrallaan yksi chat-ikkuna
- Vasemmassa laidassa näkyvät muut käynnissä olevat keskustelut ja niitä klikkaamalla voi vaihtaa näkyvissä olevan keskustelun
- Asiakkaan odotusaika kussakin keskustelussa näkyy vasemmassa laidassa keskustelun aiheen vieressä
- Otsikon alla näkyy katkelma keskustelun viimeisimmästä viestistä

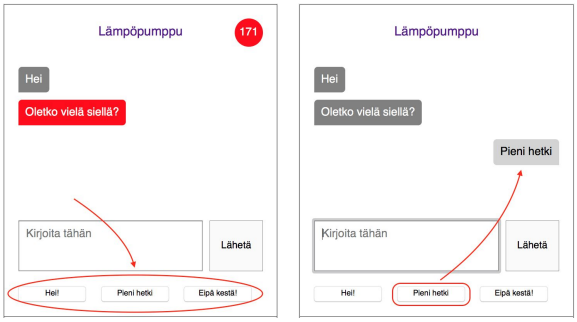


Keskustelun rakenne

- Keskustelussa on neljän tyyppisiä viestejä
  - **Tervehdys** (esim. “Hei!”)
  - **Kysymys** (esim. “Kuinka paljon astianpesukone painaa?”)
  - **Huhuilu** (“Oletko vielä siellä?”) - Jos vastaamisessa on kestänyt liian kauan
  - **Kiitos** (esim. “Kiitos vastauksista!”)

Viesteihin vastaaminen

- **Tervehdykseen**, **huhuiluun** ja **kiitokseen** vastataan vastauspainikkeilla
- Vastauspainikkeet löytyvät avoimen tekstikentän alta
- Painiketta klikkaamalla järjestelmä lähettää keskusteluun kyseisen vastauksen



Viesteihin vastaaminen

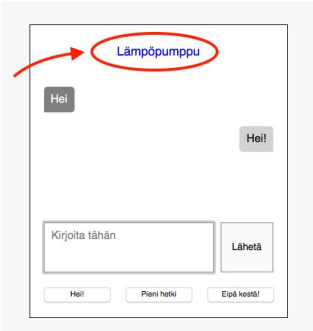
- Avoimella tekstikentällä vastataan aiheisiin liittyviin kysymyksiin
- Viesti lähetetään painamalla “Lähetä”-painiketta (ENTER ei toimi)
- Kysymyksiin vastataan mahdollisimman lyhyesti yhdellä tai enintään muutamalla sanalla, moneen vastaukseen riittää esimerkiksi numero
- Isoilla ja pienillä alkukirjaimilla ei ole väliä
- **Ei tarvitse olla kohtelias!** (se ei ole kokeen kannalta merkittävää)
- Esimerkkejä:
  - Kysymys: “Kuinka painava astianpesukone on?” - Vastaus: “60kg”
  - Kysymys: “Saako suunnistuskilpailussa käyttää piikkareita?” - Vastaus: “ei”

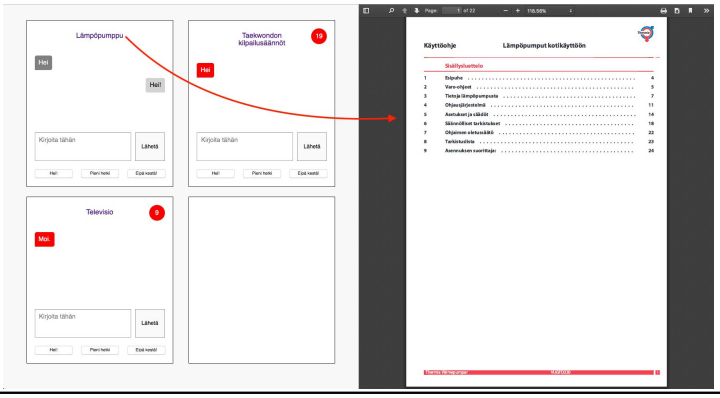
Viesteihin vastaaminen

- Tavoitteena on vastata viesteihin mahdollisimman nopeasti, jotta asiakas tuntee saavansa palvelua
  - Tähän voi käyttää “Pieni hetki” -vastauspainiketta, jos kokee että vastauksessa kestää muuten liian pitkään
  - Huomioi kuitenkin, että myös varsinainen vastaus tulisi antaa mahdollisimman nopeasti

Vastaukset kysymyksiin

- Vastaukset kysymyksiin on tarkoitus etsiä aiheeseen liittyvästä ohjekirjasta (PDF), joka aukeaa chat-ikkunoiden viereen keskustelun otsikkoa klikkaamalla





Ohjekirjan käyttö

- PDF aukeaa aina sisällysluettelosivulta
  - Jos haluaa selatessa päästä nopeasti takaisin sisällysluetteloon, voi klikata chat-keskustelun otsikkoa uudelleen
- Sisällysluettelo on interaktiivinen - kutakin otsikkoa klikkaamalla pääsee kyseiselle sivulle

Ohjekirjan käyttö

- Kysymykset on muotoiltu siten, että sisällysluettelon avulla voi päätellä, mistä vastaukset löytyvät
  - **PDF:n hakutoimintoa (tai ctrl+f) ei saa käyttää!**
  - Vinkki: laitteiden mitat ja muut ominaisuudet löytyvät usein osioista kuten “Tekniset tiedot” tai “Laitteen kuvaus”
- Pyri aina löytämään vastaus ohjekirjasta
  - Jos et kuitenkaan millään löydä vastausta, voit vastata “en tiedä” päästäksesi eteenpäin
  - Muista kuitenkin, että tavoitteena on saada mahdollisimman paljon oikeita vastauksia, ja “en tiedä” ei ole oikea vastaus

Muuta huomioitavaa

- Kokeessa testataan käyttöliittymiä, **ei käyttäjiä**!
  - Käyttäjien välisiä tuloksia ei vertailla keskenään, vaan tutkimuksen kannalta kiinnostavia ovat erot eri käyttöliittymien välillä
- Koe saattaa jossain vaiheessa tuntua intensiiviseltä
  - Tämä on osa koetta
  - Ei ole syytä ottaa paineita, pyri jatkamaan vastaamista rauhallisesti ja tehokkaasti
- Laitathan puhelimesi äänettömälle

Kysyttävää?

- Pyri esittämään ohjeisiin liittyvät kysymykset ennen kokeen aloittamista
- Mikäli kokeen aikana herää kysymyksiä, pyri esittämään kysymykset osioiden välisillä tauoilla

G Result distributions

G.1 First response time

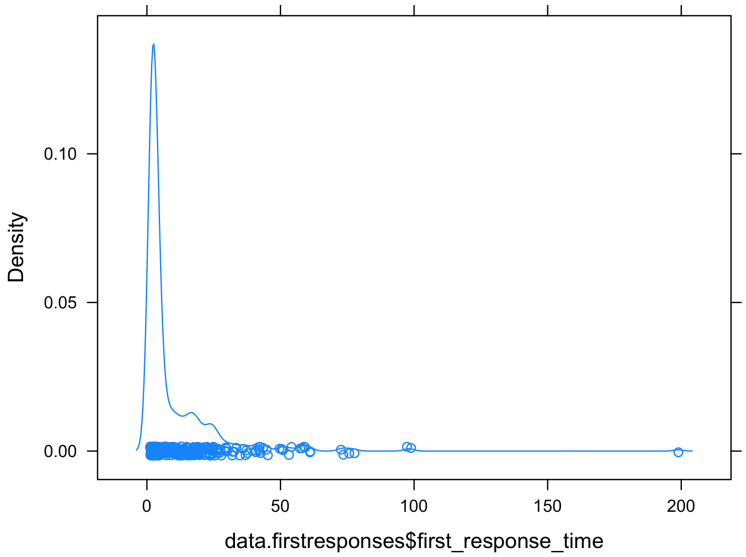


Figure G1: First response time density.

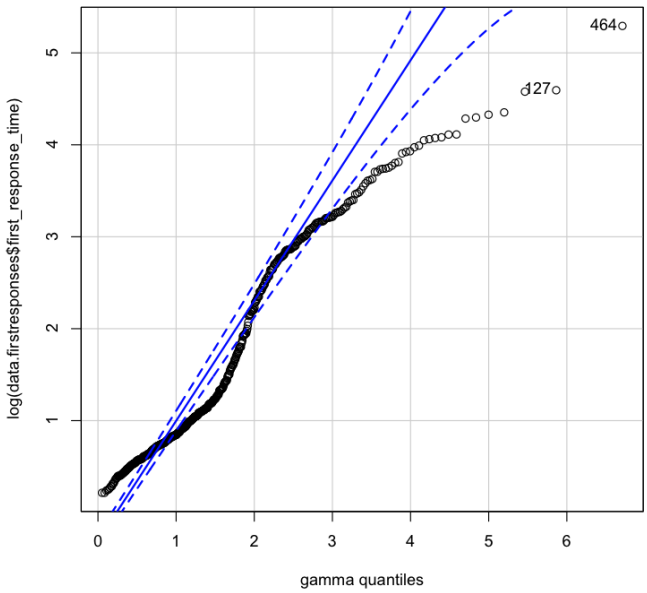


Figure G2: First response time fitted in gamma distribution.

G.2 Question response time

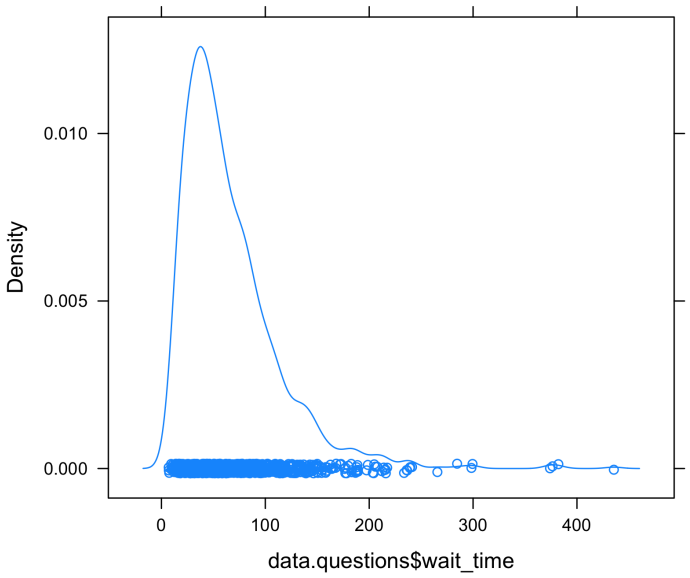


Figure G3: Question response time density.

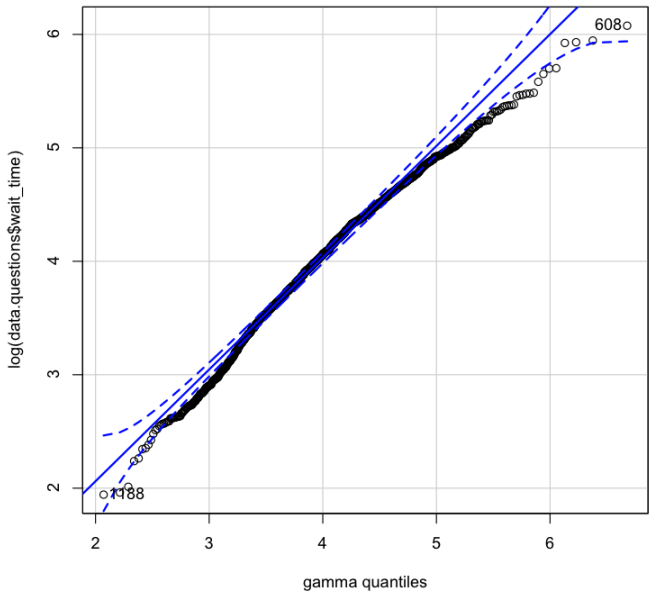


Figure G4: Question response time fitted in gamma distribution.

G.3 Accuracy

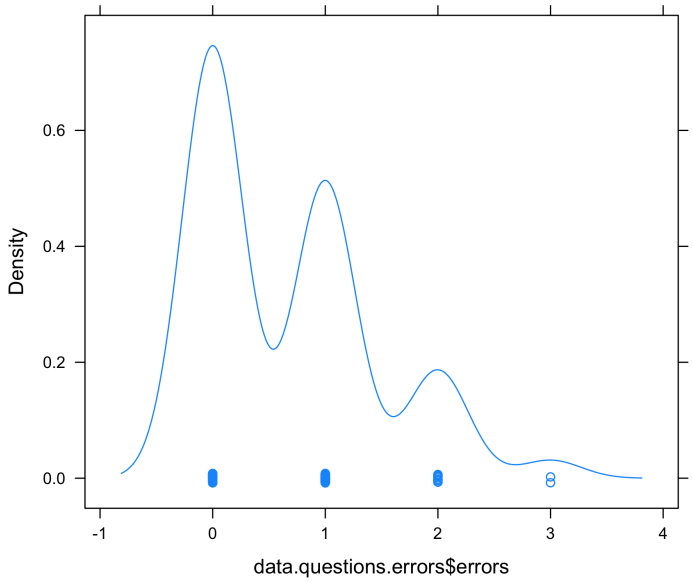


Figure G5: Number of errors density.

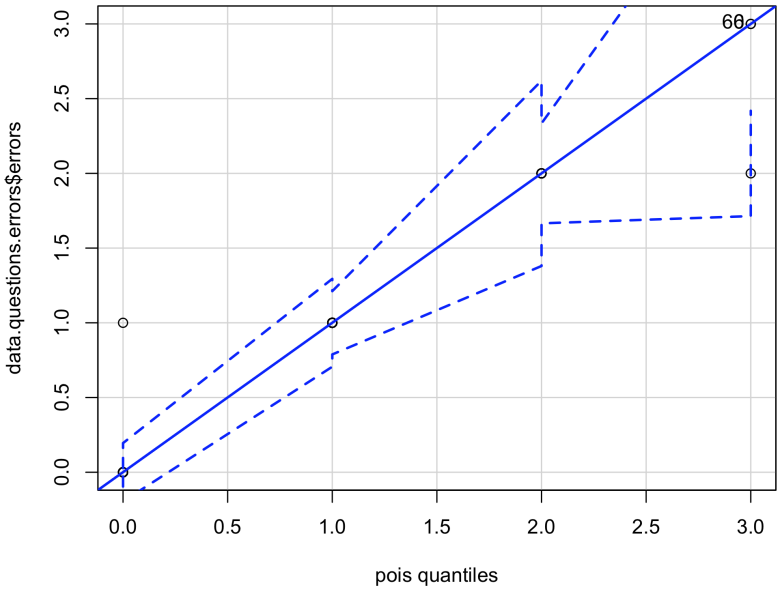


Figure G6: Number of errors fitted in poisson distribution.

G.4 Chat duration

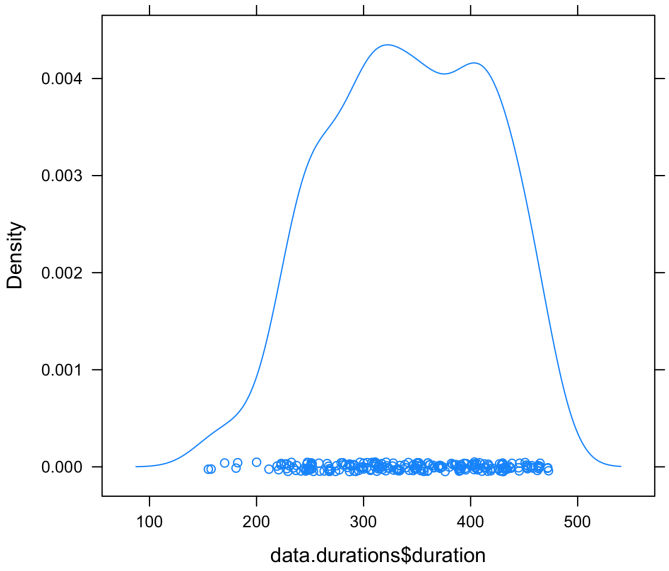


Figure G7: Chat duration density.

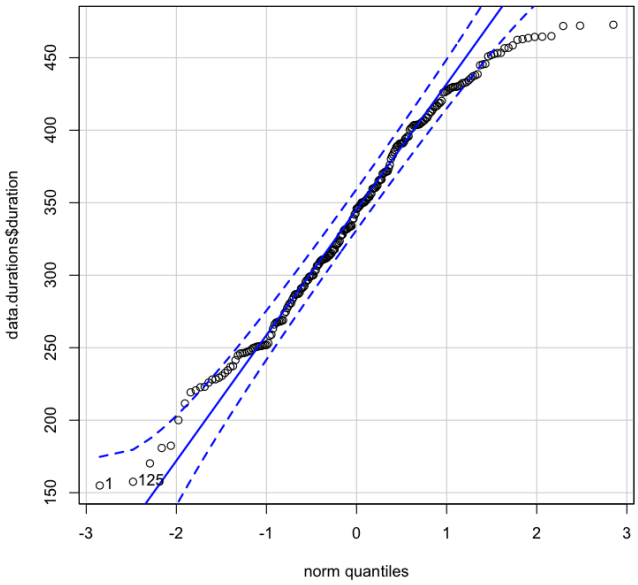


Figure G8: Chat duration fitted in normal distribution.

H Mixed model results

H.1 First response time

	Estimate	Std. error	t-value	p-value
Intercept	1.8081	0.1314	13.755	$<$ 2e-16
LayoutTabbed	0.3443	0.1285	2.731	0.00631
Chats4	-1.169	0.1248	2.760	0.00578
LayoutTabbed:Chats4	-0.1229	0.1788	-0.688	0.49170

Table H1: Generalized linear mixed model results for first response time (estimates are logarithms).

H.2 Question response time

	Estimate	Std. error	t-value	p-value
Intercept	4.00512	0.06492	61.694	$<$2e-16
LayoutTabbed	-0.05473	0.04594	-1.191	0.234
Chats4	0.37977	0.04611	8.237	$<$2e-16
LayoutTabbed:Chats4	0.06291	0.06552	0.960	0.337

Table H2: Generalized linear mixed model results for question response time (estimates are logarithms).

H.3 Accuracy

	Estimate	Std. error	t-value	p-value
Intercept	-0.51838	0.28016	-1.850	0.0643
LayoutTabbed	-0.13352	0.36596	-0.365	0.7152
Chats4	0.11779	0.34359	0.343	0.7317
LayoutTabbed:Chats4	-0.00727	0.50639	-0.014	0.9885

Table H3: Generalized linear mixed model results for number of errors (estimates are logarithms).

H.4 Chat duration

	Estimate	Std. error	*df*	*t*-value	*p*-value
Intercept	325.3801	10.1669	75.8242	32.004	< 2e-16
LayoutTabbed	-0.1309	11.8183	205.0302	-0.011	0.99117
Chats4	37.8361	12.3081	211.4006	3.074	0.00239
LayoutTabbed:Chats4	9.3928	17.8770	209.9946	0.525	0.59985

Table H4: Linear mixed model results for chat duration.